#9 attach

Marker-assisted Selection in Backcross Breeding

S.J. Openshaw

Pioneer Hi-Bred Intl. Inc., P.O. Box 1004, Johnston, IA 50131

S.G. Jarboe[1]

CIMMYT, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico

W.D. Beavis

Pioneer Hi-Bred Intl. Inc., P.O. Box 1004, Johnston, IA 50131

Abstract. **The backcross breeding procedure has been used widely to transfer simply inherited traits into elite genotypes. Genetic markers can increase the effectiveness of backcrossing by 1) increasing the probability of obtaining a suitable conversion, and 2) decreasing the time required to achieve an acceptable recovery. Simulation and field results indicated that, for a genome consisting of ten 200-cM chromosomes, basing selection on 40 or 80 markers in 50 BC individuals that carry the allele being transferred can reduce the number of backcross generations needed from about seven to three.**

The backcross breeding procedure has been used widely to transfer simply inherited traits into elite genotypes. Usually, the trait being transferred is controlled by a single gene, but highly heritable traits that are more complexly inherited have also been transferred successfully by backcrossing; for example, maturity in maize (Rinke and Sentz, 1961; Shaver, 1976). Today, backcrossing is being used to transfer genes introduced by such techniques as transformation or mutation into appropriate germplasm.

Several plant breeding textbooks give good descriptions of the backcross procedure (Allard, 1960; Fehr, 1987). A donor parent (DP) carrying a trait of interest is crossed to the recurrent parent (RP), an elite line that is lacking the trait. The F_1 is crossed back to the RP to produce the BC_1 generation. In the BC_1 and subsequent backcross generations, selected individuals carrying the gene being transferred are backcrossed to the RP. The expected proportion of DP genome is reduced by half with each generation of backcrossing. Ignoring effects of linkage to the selected DP allele being transferred, the percentage recurrent parent (%RP) genome expected in each backcross generation is calculated as:

$$\%RP = 100\,[1 - (0.5)^{n+1}]$$

where n is the number of backcrosses.

Backcrossing of selected plants to the RP can be repeated each cycle until a line is obtained that is essentially a version of the RP that includes the introgressed allele. After six backcrosses, the expected recovery is >99% (Table 1).

Until recently, discussions of the recovery of the RP genome during backcrossing have emphasized the expected values for %RP shown in Table 1, and have largely ignored the genetic variation for %RP that exists around the expected mean. With the development of genetic markers capable of providing good genome coverage, there has been interest in taking advantage of that variation to increase the efficiency of backcrossing.

Selection for RP marker alleles can increase greatly the effectiveness of backcross programs by allowing the breeder to: 1) select backcross plants that have a higher proportion of RP genome, and 2) select backcross individuals that are better conversions near a mapped donor allele being transferred (i.e., select for less linkage drag). Expressed in practical terms, using genetic markers to assist backcrossing can 1) increase the probability of obtaining a suitable conversion, and 2) decrease the time required to achieve an acceptable recovery.

Issues to consider when planning a marker-assisted backcross program include 1) the time advantage of using markers to assist backcrossing, 2) the number of markers needed, and 3) the number of genotypes to evaluate. In this report, we use results from previous literature, computer simulation, and empirical studies to provide some guidelines.

Table 1. *Expected recovery of recurrent parent (RP) genome during backcrossing, assuming no linkage to the gene being transferred.*

Generation	%RP
F_1	50.0000
BC_1	75.0000
BC_2	87.5000
BC_3	93.7500
BC_4	96.8750
BC_5	98.4375
BC_6	99.2188
BC_7	99.6094

[1]Formerly with Purdue University, West Lafayette, Ind.

Appendix A

Materials and methods

The maize genome was the model for the simulation. The simulated genome contained ten 200-cM chromosomes. Simulation of crossing over was based on a Poisson distribution with a mean of 2.0 ($\lambda = 2$) (Hanson, 1959), which, on average, generated one cross over for every 100-cM length. The simulations reported here assume no interference. Codominant genetic markers were evenly distributed in the genome and sites of the donor gene were randomly assigned to genome locations. Simulations were conducted with the following parameters:

Number of progeny: 100 or 500.
Backcross generations: BC_1, BC_2, and BC_3.
Number of markers: 20, 40, 80, or 100.
Number selected to form the next BC generation: 1 or 5.

Selection was based on 1) presence of the donor allele and 2) high %RP). %RP was calculated as the average of the (one or five) selected individuals. Values presented are the mean of 50 simulations.

Results

In the computer simulation study, all methods modeled greatly increased the speed of recovering the RP genome compared to the expected recovery with no marker-assisted selection (compare Tables 1 and 2). At least 80 markers were required to recover 99% of the RP genome in just three BC generations (Table 2). Use of at least 80 markers and 500 progeny allowed recovery of 98% RP in just two BC generations. Response to selection was diminished only slightly by spreading the effort over five selections. Using markers, the number of backcross generations needed to convert an inbred is reduced from about seven to three.

By the BC_1 generation, there appears to be no practical advantage to using 500 vs. 100 individuals. If the presence of the donor trait in the backcross individuals can be ascertained before markers are genotyped, then only half the number of individuals indicated in the tables will need to be analyzed.

When a small number of markers are used, they quickly became non-informative; i.e., selection causes the marker loci to became fixed for the RP type before the rest of the genome is fully converted (Table 3; Hospital et al., 1992). This situation was most prominent in the larger populations, where a higher selection intensity placed more selection pressure upon the marker loci. Accordingly, it is of interest to consider how closely the estimation of %RP based on markers reflects the actual genome composition. The combination of estimation of %RP based on fewer markers and subsequent selection tends to bias the estimates upward (compare Tables 2 and 3).

The results from the simulation compare well with real field data. In a typical example, 50 BC_1 plants carrying the gene being transferred were genotyped at 83 polymorphic RFLP loci (note that this corresponds to a population size of 100 unselected plants in Tables 2 and 3). The five best BC_1 recoveries had estimated %RP values of 85.9%, 82.7%, 82.0%, 81.4%, and 81.2%. After evaluating 10 BC_2 plants from each selected BC_1, the best BC_2 recovery had an estimated %RP of 94.6%.

Discussion

The simulations (Table 2; Hospital et al., 1992) and our experience indicate that four markers per 200-cM chromosome is adequate to greatly increase the effectiveness of selection in the BC_1. However, using only four markers per 200 cM will likely make it very difficult to map the location of the gene of interest. Adequate summarization of the data is an important

Table 2. *Percent recurrent parent genome during marker-assisted backcrossing.*

	100 Progeny				500 Progeny			
	No. markers				No. markers			
Generation	20	40	80	100	20	40	80	100
				One selected				
BC_1	84.5	84.5	84.2	88.0	89.9	90.7	90.2	90.5
BC_2	95.0	95.2	95.8	97.2	96.5	97.7	98.5	98.6
BC_3	97.4	97.6	98.9	99.2	97.7	98.3	99.4	99.5
				Five selected				
BC_1	82.9	85.1	84.9	84.7	87.7	88.1	88.9	88.9
BC_2	93.7	95.0	95.8	95.7	95.5	96.8	97.8	97.9
BC_3	97.1	98.3	98.8	98.9	97.3	98.5	99.3	99.3

Table 3. *Estimates of percent recurrent parent genome, based on marker loci.*

	100 Progeny				500 Progeny			
	No. markers				No. markers			
Generation	20	40	80	100	20	40	80	100
				One selected				
BC_2	98.7	97.8	95.6	97.2	100.0	99.1	98.6	98.0
BC_3	100.0	99.8	99.3	99.5	100.0	100.0	99.9	98.2
				Five selected				
BC_2	96.4	96.5	96.2	95.8	100.0	98.5	98.3	98.2
BC_3	99.9	99.8	99.3	99.1	100.0	100.0	99.9	99.8

part of a marker-assisted backcross program. Ideally, the markers used can supply data that can be represented as alleles of loci with known map position. Estimation of %RP, mapping the position of the locus of interest, and graphical display of the results (Young and Tanksley, 1989) are all useful in understanding and controlling the specific backcross experiment being conducted.

It appears that, with the use of genetic markers, the portion of the RP genome that is not linked to the allele being transferred can be recovered quickly and with confidence. The recovery of RP will be slower on the chromosome carrying the gene of interest. A considerable amount of linkage drag is expected to accompany selection for the DP allele in a backcross program. For a locus located in the middle of a 200-cM chromosome, the length of the DP chromosome segment accompanying selection is expected to be 126, 63, and 28 cM in the BC_1, BC_2, and BC_3 generations, respectively (Hanson, 1959; Naveira and Barbadilla, 1992). Our observations support the recommendation of Hospital et al. (1992) that preference be given to the selection for recombinants proximal to the allele of interest, but that selection for recovery of the RP elsewhere in the genome also be considered. This two-stage selection can probably be done quite effectively ad hoc by the breeder once the data is adequately summarized; however, Hospital et al. suggest ways to incorporate the two criteria into a selection index such that each component of selection is assured appropriate weighting.

Use of genetic markers can greatly increase the effectiveness of backcrossing, and they should be used in any serious backcrossing program if resources are available to the breeder.

Literature Cited

Allard, R.W. 1960. Principles of plant breeding. Wiley, New York.

Fehr, W.F. 1987. Principles of cultivar development v.1. Theory and technique. Macmillan, New York.

Hanson, W.D. 1959. Early generation analysis of length of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. Genetics 44:843–847.

Hospital, F., C. Chevalet, and P. Mulsant. 1992. Using markers in gene introgression breeding programs. Genetics 132:1199–1210.

Rinke, E.H and J.C. Sanz. 1961. Moving corn-belt germplasm northward. Ann. Hybrid Corn Industry Conf. 16:53–56.

Shaver, D.L. 1976. Conversions for earliness in maize inbreds. Maize Genet. Coop. Newsltr. 50:20–23.

Young, N.D. and S.D. Tanksley. 1989. Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theo. Applied Genet. 77: 95–101.

2 MAIZE - THE PLANT AND ITS PARTS

R. Scott Poethig
Department of Agronomy, Curtis Hall
University of Missouri
Columbia, MO 65211

One of the greatest deterrents to an appreciation of plant morphology is the terminology used to describe various plant parts. This problem is compounded in the case of maize because of its relatively unusual structure. We all learn that plants have a vegetative body composed of stems, leaves and roots, and that flowers contain sepals, petals, pistils and stamens. Maize, however, has at least three kinds of leaves, two kinds of stems, two kinds of roots, and two kinds of flowers in which glumes, lemmas and paleas take the place of sepals and petals. Fortunately, these parts are arranged in a relatively simple fashion, so the task of mastering maize morphology is not as difficult as it might seem. In this article we will identify some of the most important parts of the maize plant and describe their organization. More detailed descriptions of the developmental morphology of maize have been provided by a number of investigators. Kiesselbach (1949, reprinted 1980) gives a good general picture of maize structure and development. The external morphology and the histology of the vegetative and reproductive shoots have been studied by Bonnett (1948, 1953), Sharman (1942) and Abbe and co-workers (Abbe and Phinney, 1951; Abbe et al., 1951), while the most comprehensive descriptions of the embryogeny are those of Randolph (1936) and Abbe and Stein (1954). A summary of the histology of the corn plant, written by Sass in 1955, has been reprinted in the recent edition of Corn and Corn Improvement (1976).

The organization of the plant body: Maize is a member of the grass family, the Gramineae, and as in all grasses, most of the plant body is leaf tissue (Fig. 1a). To appreciate the general organization of the maize plant it is helpful, therefore, to see it in a leaf-less state (Fig. 1b). Stripped naked, the maize plant is not very impressive. Its main stem, or culm, is a slender, segmented shaft similar to a stalk of bamboo or sugarcane. The enlarged joints along the stem, the nodes, mark the points of leaf attachment; the stem segment between nodes is called the internode. Each node bears a single leaf in a position opposite that of the neighboring leaf, giving the plant two vertical rows of leaves in a single plane (Fig. 1a; 2). This so-called distichous phyllotaxy is typical of all leaf-like appendages, wherever they occur on the plant.

Maize has unisexual, rather than bisexual flowers. Male (staminate) flowers are located at the apical tip of the main stem in the tassel, a branched inflorescence. Female (pistillate) flowers are found in one to several compact ears, located on the ends of short branches near the middle of the stem (Fig. 1b; 2).

This partitioning of male and female flowers in separate structures distinguishes maize from other cereals and is one of the principal reasons that its genetics has been so conveniently explored. Making controlled pollinations in maize requires little more effort than that involved in placing a bag over the tassel and ear shoot. To perform a controlled pollination in rice, wheat, barley and other cereals, it is necessary to emasculate each

flower used as a female parent, an especially tedious job when each flower yields only one seed.

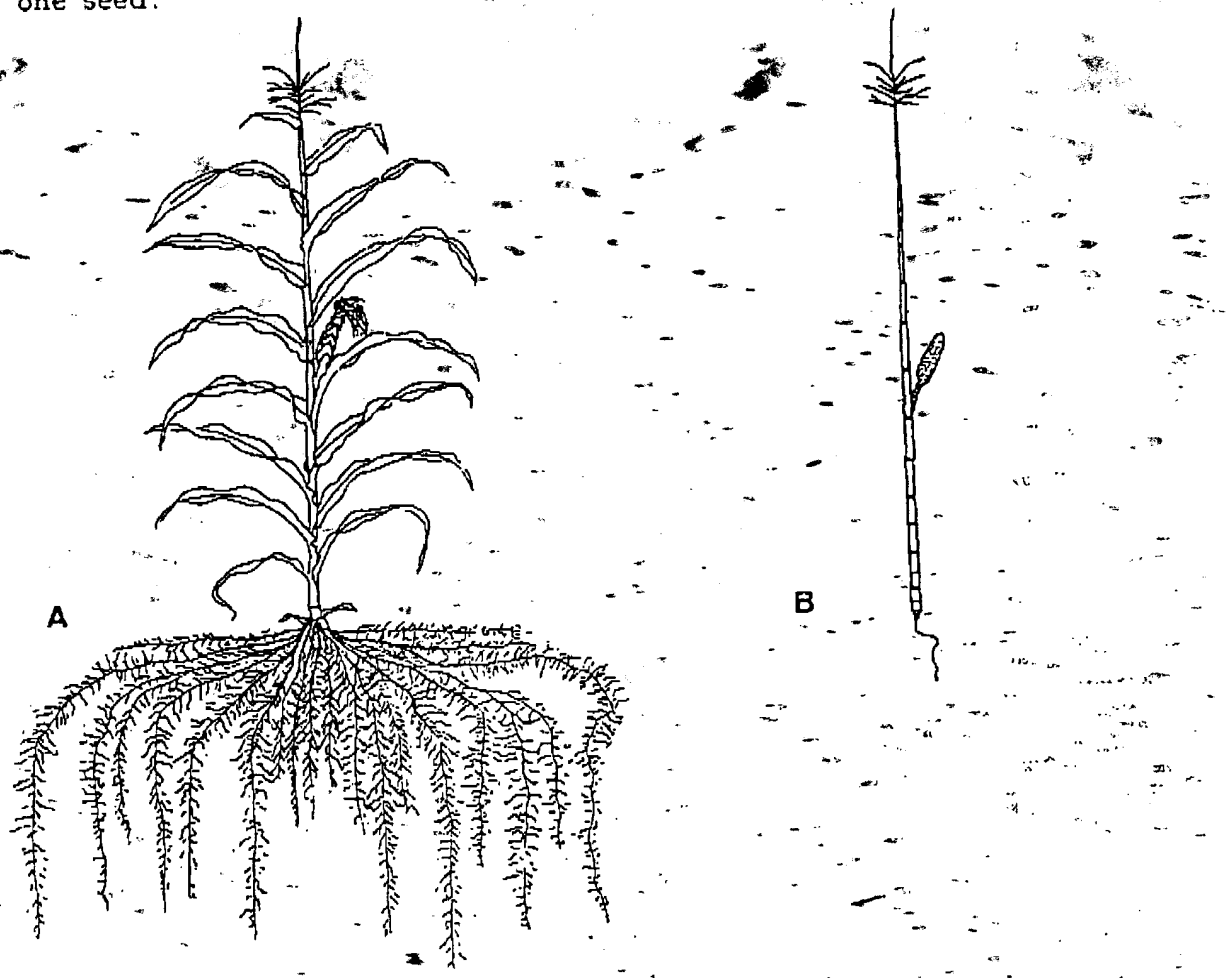

Figure 1. a) Mature maize plant (after Kiesselbach, 1949). b) Mature maize plant drawn without leaves and adventitious roots. The apical end of the main stem (culm) terminates in the tassel, while the basal end terminates in the primary root (radicle). The ear shoot arises from an internode near the center of the culm.

Maize also differs from closely related species in that it has relatively few branches. Only the lower 10 to 12 internodes of the stem produce branch primordia, and most of these remain suppressed. Above-ground primordia develop into ear shoots, while those located at subterranean internodes develop into tillers--branches identical in structure to the main stem. Commercial hybrids (except sweet corns) generally tiller very little, and typically produce a single viable ear shoot. In contrast, some "varieties" may have several large tillers and may produce 2 ears on the main stem and some ears on tillers.

The stem: During the first four weeks after germination, the growing point of the stem lays down all the nodes and internodes of the plant and then differentiates into a tassel. At the time of tassel formation the stem is not more than 3-4 inches tall, even though the plant may be 3-4 feet in

height (Fig. 3). Subsequently, the stem begins to elongate rapidly, with most of the growth occurring at the base of the internodes. The lowermost 6-8 internodes do not participate in this growth, however, and remain below ground where they produce the root system and tillers. These subterranean internodes taper sharply towards the base of the stem, forming a distinctive region, the crown (Fig. 1b). The stem is thickest a few inches above ground, and tapers gradually towards the tassel. All the internodes from the top ear downward have a distinct groove associated with the axillary bud at the base of the internodes; internodes above the ear lack axillary buds and are smoothly cylindrical.

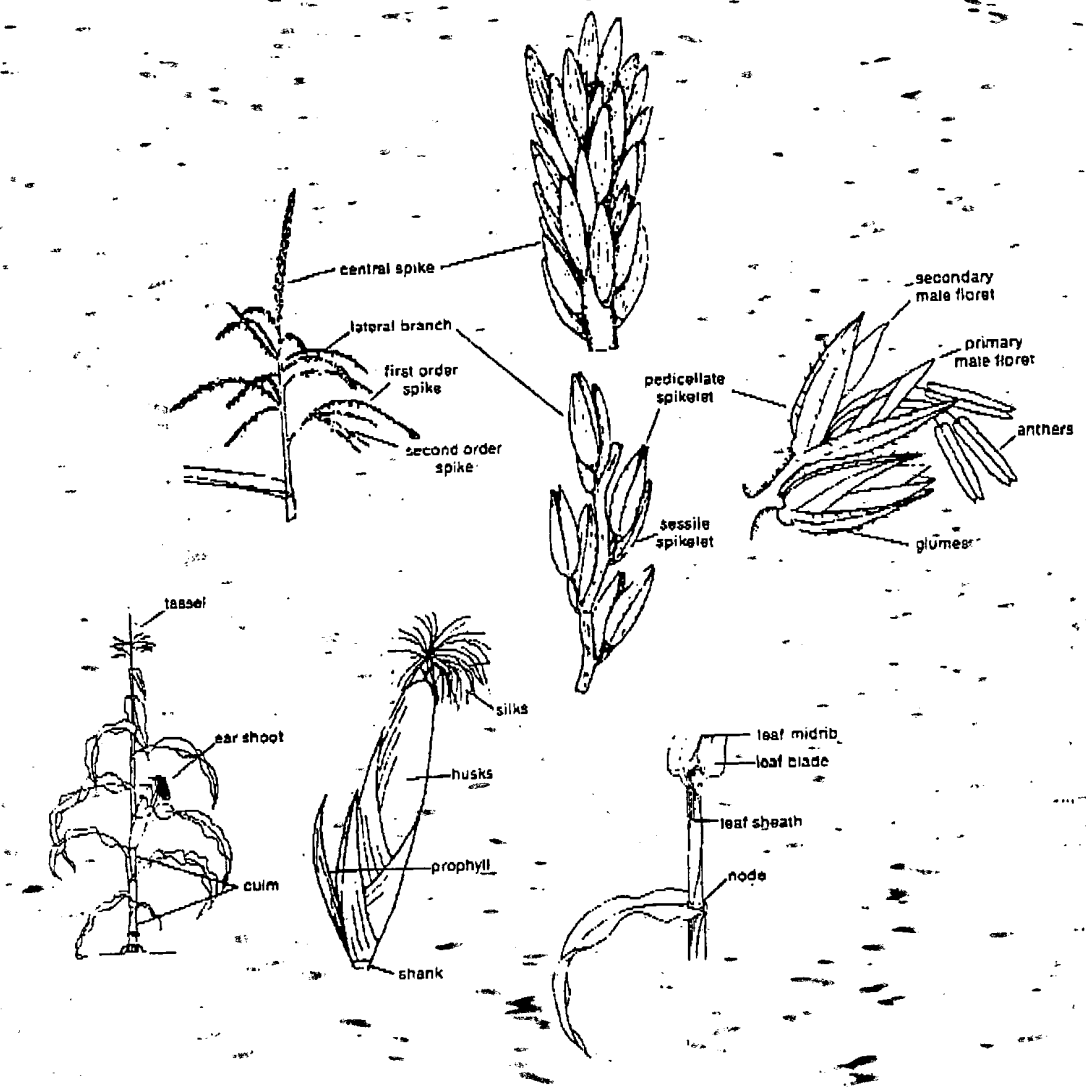

Figure 2. The major parts of the maize plant. Drawings in part from P. Weatherwax in Corn and Corn Improvement, 1955, and E. D. Styles et al. in Can. J. Genet. Cytol. 15:59, 1973; figure assembled by M. M. Johri and E. H. Coe.

The stem of an ear shoot, called the shank (Fig. 2), differs from the main stem in being relatively short in most strains. In addition, the internodes of the shank are variable in number, irregular in shape and size, and tend to have a crinkled rather than smooth surface. Secondary ear shoots commonly occur on the shank of several types of maize, but are rare in most commercial strains unless fertilization of the apical ear is prevented.

Figure 3. A four week old plant (approximately 3 feet tall) in which the stem apex has differentiated into a tassel. As shown on the right, the stem is still relatively short at this stage.

The tassel: The tassel, located at the top of the culm, consists of a series of large branches (spikes) covered with numerous, small flower-bearing branches (spikelets: Fig. 2) Each branch point on a spike bears two spikelets, one on a long stem (pedicellate), the other on a short-stem (sessile) (Fig. 4a). Each of these spikelets, in turn, produces two functional florets. Although tassel florets contain both stamens and a pistil, the pistil normally degenerates soon after it is initiated, making the floret functionally male. However, pistils will develop at the base of the tassel under some environmental and physiological conditions, and are quite common on tillers.

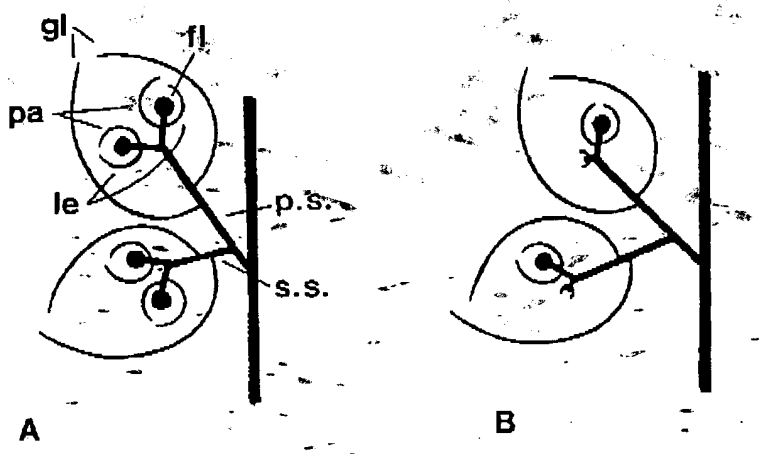

Figure 4. Schematic drawing of a pair of tassel spikelets (A) and a pair of ear spikelets (B). Note that the lower floret in the ear spikelet aborts early in development. p.s. - pedicellate spikelet; s.s. - sessile spikelet; gl - glumes; le - lemma; pa - palea; fl - floret.

Surrounding both florets on a spikelet are 2 leaf-like scales called glumes (Fig. 2; 4a). Within the glumes, each floret is individually enclosed in another pair of scales, one located adjacent to the glume (the lemma), the other located between the two florets (the palea) (Fig. 4a). At anthesis, these scales are forced apart by the swelling of conical structures (lodicules) at the base of the 3 stamens, and the filamentous base of the stamens elongates, forcing the anthers out of the flower (Fig. 2). As they dangle downwards, the anthers shed pollen from openings at their tip.

Pollen grains are the multicellular products of the haploid microspores that result from the meiosis of a microspore mother cell (microsporocyte). Meiosis takes place in the anther before the tassel emerges from the leaf sheaths. After meiosis, the 4 resulting haploid microspores separate from each other, and each forms a thick wall. Shortly before shedding, each microspore undergoes two mitotic divisions. The first division is asymmetric, and produces a relatively large vegetative cell and a smaller generative cell. In the second division, the generative cell divides to form two sperm cells.

The ear: The ear is morphologically similar to the tassel, although this resemblance is obscured by differences in the relative size of their parts. The crucial difference between them is, of course, that the tassel contains male flowers, and the ear bears female ones. This difference is due simply to the fact that during the formation of an ear floret, stamen primordia are arrested at an early stage in their development, while the pistil develops fully. Each functional ear floret has a single ovary, which terminates in an elongated style, or silk (Fig. 5). Within the ovary is a single embryo sac. The embryo sac is the product of one of the four haploid cells resulting from the meiosis of the megaspore mother cell. While its three sister cells degenerate, the nucleus of this cell divides three times to produce 8 haploid nuclei within a common cytoplasm (the embryo sac). Two of these nuclei (polar nuclei) migrate to the center of the embryo sac, where they become closely associated. The three nuclei remaining at the base of the embryo sac

subsequently undergo cellularization to form the egg cell and two synergids, while the 3 nuclei at the tip of the embryo sac proliferate to form 24-48 antipodal cells.

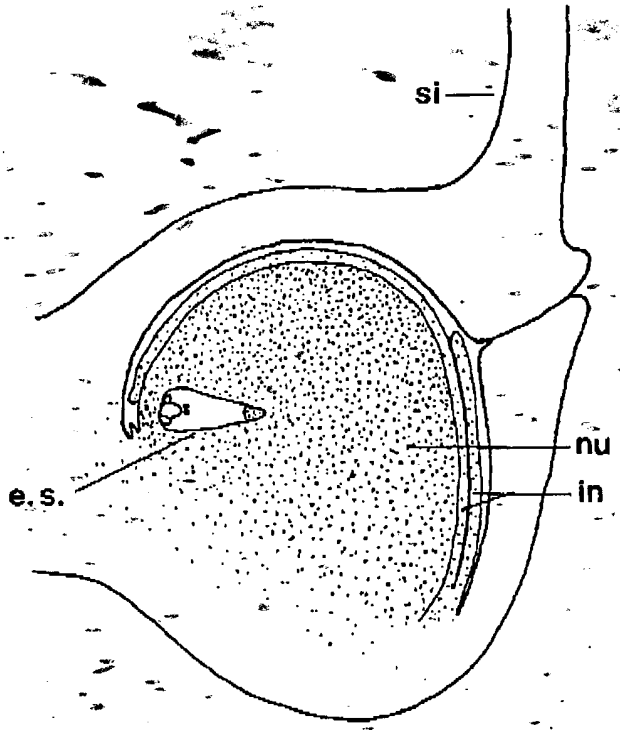

Figure 5. Radial longitudinal section of an ovary with an unfertilized embryo sac (after Randolph, 1936). Upon fertilization, the nucellus is digested by the expanding embryo sac and the tissue surrounding the nucellus is transformed into the pericarp. si - silk; e.s. - embryo sac; nu - nucellus; in - integuments.

The ear also differs from the tassel in that it has no major lateral branches. Its thick, lignified axis, the cob, is homologous to the central spike of the tassel. As in the tassel, ear spikelets come in pairs, but in the ear they are equal in size and only one of the florets in each spikelet is functional (Fig. 4b). An ear therefore has an even number of parallel rows of equally sized kernels equal to the number of spikelets on the cob. The number of rows (or ranks) of kernels ranges from 4 to 30.

The glumes, lemmas and paleas of the ear spikelets are readily visible in an unfertilized ear, but are soon obscured by the enlargement of the ovary after fertilization. In a mature ear these structures are represented by the chaff that adheres to the cob and the base of the kernel after it is shelled.

The leaf: Maize produces three kinds of vegetative leaves: foliar leaves, husk leaves and prophylls. A foliar leaf is located at each of the nodes on the main stem, husk leaves are located on the shank of the ear shoot, and prophylls are found at the base of the shank between the ear shoot and the stem (Fig. 2).

The foliar leaf has two distinct parts--the blade, a flat portion extending away from the stem, and the sheath, a basal part that wraps tightly around the stem (Fig. 2). Internally, the blade consists of a spongy network of cells traversed by a series of parallel, longitudinal veins. This flexible lamina is supported by the midrib, a thickened, translucent structure located in the center of the leaf. The sheath is thicker and more rigid than the blade, possesses fewer longitudinal veins, and lacks a prominent midrib. The sheath completely encircles the internode above the node to which it is attached and may extend the entire length of that internode. During the early development of the plant, the leaf sheaths provide most of the mechanical support necessary to keep the stem upright. At the boundary between the blade and the sheath there is a distinct hinge of translucent tissue. In this region, the leaf blade and leaf sheath narrow sharply, forming an indentation in the leaf margin. The wedge of translucent tissue adjacent to this indentation is known as the auricle. The ligule is the thin collar of filmy tissue located on the inside of the hinge.

The husk leaves surrounding the ear are usually considered modified leaf sheaths, with vestiges of the blade portions occasionally present. In some strains husk leaves develop a prominent ligule and leaf blade. In contrast to the leaf sheath, husk leaves are relatively thin and flat. Each husk leaf is attached to a unique node on the shank, and all but a few upper ones are arranged distichously.

Located between an ear shoot and the stem, the prophyll looks superficially like a husk leaf, but is distinguished by having two keels (midribs) and a split apex. These features suggest that the prophyll arose evolutionarily from the fusion of two foliar leaves. The homology of the prophyll is still controversial, however. Galinat (1959), for example, considers the prophyll one of the basic units of maize morphology, the others being the internode, leaf and axillary bud.

The root: More is known about the growth, cell biology, physiology and anatomy of the primary maize root, or radicle, than perhaps any other organ of the plant. Its histological structure, described by Sass (1976) and Kiesselbach (1949), is typical of roots in general. The apex of the root is sheathed in a loose network of root cap cells. Immediately behind the apex is a zone of cell division and elongation, beyond which root hairs are initiated. Larger lateral roots arise at varying points behind the zone of root hair formation. Cell division is restricted to the apical 3 mm of the root, and occurs at a maximal rate 1.25 mm behind the apex. The zone of elongation extends 8 mm behind the apex, the rate of elongation being maximal 4 mm from the tip (Erickson and Sax, 1956). Those interested in using the root for physiological or cell cycle studies should consult Silk and Erickson (1979; 1980) and Green (1976) for an analysis of the growth parameters that must be taken into consideration in such studies.

The primary root represents the basal end of the plant axis, which in maize and other grasses contributes relatively little to the ultimate root system (compare Fig. 1a and b). Most of the root system consists of adventitious roots produced by the basal-most internodes of the stem. The primordia of a few adventitious roots are normally present in the embryo, and these emerge soon after germination. New root primordia ar subsequently initiated at the base of all subterranean internodes, and also appear

at 2 or 3 above-ground internodes after the stem has elongated. Subterranean adventitious roots are sometimes called crown roots, while those initiated above ground are known as brace roots.

Adventitious roots grow horizontally for several feet before turning downwards. As a result, the root system of a single plant often covers a region 6-8 feet in diameter, while the depth of the root system may be as much as 6 feet. As it grows, the root branches profusely in the region behind the apex, forming both secondary roots and unicellular root hairs. The total length of root system of a mature plant has been estimated to be 6 miles.

The kernel: The events surrounding the process of fertilization have been described by Miller (1919), Kiesselbach (1949) and Pfahler (1975); unfortunately, ultrastructural information about this phenomenon is still unavailable.

The silk is receptive to pollen along its entire length. Within 5 minutes after a pollen grain lands on a silk it sends out a tube which penetrates the silk and grows downward towards the ovary. During this process the vegetative nucleus and the two sperm cells migrate to the tip of the pollen tube where they remain throughout its growth. Upon reaching the embryo sac, 12 to 24 hours after germination, the end of the pollen tube bursts, releasing the two sperm. One sperm nucleus fuses with the two polar nuclei in the center of the embryo sac to form a triploid cell that gives rise to the endosperm. The other sperm nucleus fuses with the egg nucleus to form the zygote. As often as 2% of the time the polar nuclei and the egg nucleus are fertilized by sperm from different pollen grains, with the extra sperm nuclei being somehow lost (Sarkar and Coe, 1971). This phenomenon, called heterofertilization, can lead to a non-correspondence between the genotype of the endosperm and embryo when the male parent is heterozygous.

The development of the kernel following fertilization has been described in detail by Randolph (1936). We will only note here that this process takes 40-50 days and is accompanied by a 1400-fold increase in the volume of the embryo sac. The growth of the embryo and the accumulation of food reserves in the endosperm is completed by about day 40, and the remaining 10-20 days is spent maturing and drying.

A mature kernel has three major parts: the pericarp, endosperm and embryo (Fig. 6). The pericarp, the tough transparent outer layer of the kernel, is derived from the ovary wall and is therefore genetically identical to the maternal parent. The endosperm and embryo represent the next generation.

The endosperm makes up about 85% of the weight of the kernel and is the food source for the embryo for several days after it germinates. This food takes the form of intracellular starch grains and protein bodies, and is concentrated to varying degrees in different parts of the endosperm (Duvick, 1961). In flint-type kernels the concentration of starch and protein bodies is higher around the periphery of the endosperm than in the center, giving the endosperm a hard, corneous external layer, and a soft, granular center. In dent kernels, the granular tissue extends to the crown of the endosperm so that it collapses upon drying and produces a distinct indentation. These two traits are polygenic in their inheritance and are

characteristic of specific races of maize. Other common endosperm traits, such as sugary, floury or shrunken, are single gene mutations and can exist in either a flint or dent background.

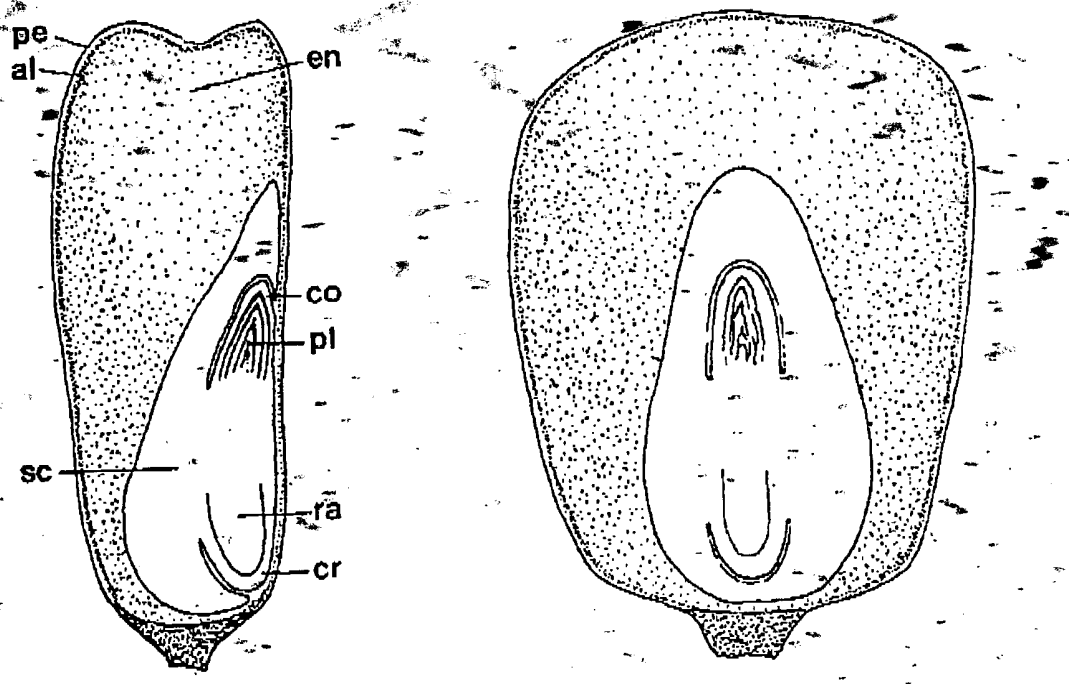

Figure 6. Longitudinal sectors of a mature dent kernel, taken perpendicular (left) and parallel (right) to the upper face of the kernel (after Kiesselbach, 1949). pe - pericarp; en - endosperm; al - aleurone; sc - scutellum; co - coleoptile; pl - plumule; ra - radicle; cr - coleorhiza.

Much of our understanding of gene action in maize is based on the analysis of genes affecting the pigmentation of the external layer of the endosperm, the aleurone. This specialized single cell layer is the only part of the endosperm capable of becoming intensely pigmented. Internal endosperm cells may be either yellow or white.

The embryo is located on the broad side of the kernel facing the upper end of the ear, beneath a thin layer of endosperm cells. Most of the tissue in the embryo is part of the scutellum, a spade-like structure concerned with digesting and transmitting to the germinating seedling the nutrients stored in the endosperm. The shoot and root axis are recessed in the outer face of the scutellum. In a mature kernel, the shoot (plumule) has 5 to 6 leaf primordia that are arrested at successive stages of development (Abbe and Stein, 1954). Surrounding the shoot is a cylindrical structure called the coleoptile. Upon germination, the coleoptile elongates until it is above ground and is then ruptured by the more rapid expansion of the rolled leaves within it. The root is enclosed in a sheath of tissue called the coleorhiza. Unlike the coleoptile, the coleorhiza does not elongate very much, and gives way to the radicle as soon as it emerges from the seed.

References

Abbe, E. C. and B. O. Phinney. 1951. The growth of the shoot apex in maize: external features. Amer. J. Bot. 38:737-744.

Abbe, E. C., B. O. Phinney and D. F. Baer. 1951. The growth of the shoot apex in maize: internal features. Amer. J. Bot. 38:744-751.

Abbe, E. C. and O. L. Stein. 1954. The growth of the shoot apex in maize: embryogeny. Amer. J. Bot. 41:285-293.

Bonnett, O. T. 1948. Ear and tassel development in maize. Missouri Bot. Gard. Ann. 35:269-287.

Bonnett, O. T. 1953. Developmental morphology of the vegetative and floral shoots of maize. Bull. 568, Agric. Exp. Sta., U. of Illinois.

Duvick, D. N. 1961. Protein granules of maize endosperm cells. Cereal Chem. 38:374-385.

Erickson, R. O. and K. B. Sax. 1956. Elemental growth rate of the primary root of Zea mays. Proc. Amer. Phil. Soc. 100:487-498.

Galinat, W. C. 1959. The phytomer in relation to floral homologies in the American Maydeae. Bot. Mus. Leaflets, Harvard U. 19:1-32.

Green, P. B. 1976. Growth and cell pattern formation on an axis: critique of concepts, terminology and modes of study. Bot. Gaz. 137:187-202.

Kiesselbach, T. A. 1949. The structure and reproduction of corn. Res. Bull. 161, Agric. Exp. Sta., U. of Nebraska College of Agric. Reprinted 1980, U. Nebraska Press.

Miller, E. C. 1919. Development of the pistillate spikelet and fertilization in Zea mays L. J. Agric. Res. 18:255-265, with 14 plates.

Pfahler, P. L. 1978. Biology of the maize male gametophyte. In: Maize Breeding and Genetics D. B. Walden, ed., John Wiley and Sons, Inc.

Randolph, L. F. 1936. Developmental morphology of the caryopsis in maize. J. Agric. Res. 53:881-916.

Sarkar, K. R. and E. H. Coe, Jr. 1971. Analysis of events leading to heterofertilization in maize. J. Hered. 62:118-120.

Sass, J. E. 1976. Morphology. In: Corn and Corn Improvement G. F. Sprague, ed., Amer. Soc. Agronomy, Madison.

Sharman, B. C. 1942. Developmental anatomy of the shoot of Zea mays L. Ann. Bot. N. S. 6:245-282.

Silk, W. K. and R. O. Erickson. 1979. Kinematics of plant growth. J. Theor. Biol. 76:481-501.

Silk, W. K. and R. O. Erickson. 1980. Local biosynthetic rates of cytoplasmic constituents in growing tissue. J. Theor. Biol. 83:701-703.

Attorney Dock t No. 1328

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: Joseph Kevin Gogerty Date: February 28, 2003

Serial No.: 09/758,867 Group Art Unit: 1638

Filed: January 11, 2001 Examiner: David T. Fox

For: "INBRED MAIZE LINE PH7CH"

Assistant Commissioner for Patents
Washington, D.C. 20231

RULE 132 DECLARATION
OF
DR. STEPHEN SMITH

Sir:

I, Stephen Smith, PhD., do hereby declare and say as follows:

1. I am skilled in the art of the field of the invention. I have a Ph.D. in Biochemical Systematics and Taxonomy of Maize and its Wild Relatives from Birmingham University. I have a M.Sc. in the Conservation and Utilization of Plant Genetic Resources from Birmingham University. I have a Bachelor of Science degree in Plant Sciences from London University. Since 1977 I have been engaged in the development, study and application of molecular markers to genetics, measuring genetic diversity and tracking pedigrees. I commenced this work at North Carolina State University as a post-doctoral research fellow. I have continued my engagement in these studies during my employment by Pioneer Hi-Bred from 1980 until the present. These studies have resulted in numerous scientific articles that have appeared in peer-reviewed scientific literature.

2. I have read and understood the Office Action in the above case dated October 30, 2002. This declaration is in response to the Examiner's rejection under, 35 U.S.C. § 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

3. I have conducted an analysis of Simple Sequence Repeat, SSR, marker data for base inbred PH7CH and a backcross conversion of PH7CH. The trait backcrossed into the backcross conversion of PH7CH was waxy starch.

Appendix B

4. The SSR data for 457 base inbreds and 103 backcross conversion inbreds, including PH7CH and the backcross conversion was used in the analysis. The number of SSR markers for each inbred used in the analysis was between 15 and 87 (mean of 82). The analysis was done as specified in the publication by Berry et al. ("Assessing Probability of Ancestry Using Simple Sequence Repeat Profiles: Applications to Maize Hybrids and Inbreds" Genetics 161:813-824, 2002); with modification as described in Berry et al., (2003); Assessing Probability of Ancestry Using SSR Profiles: Application to maize inbred lines and soybean varieties. Genetics (in review); a copy of which is attached hereto.

5. The results of the analysis indicated that through the use of SSR markers PH7CH was identified to be the recurrent parent of the backcross conversion of PH7CH over all the other inbreds in the data set. The probability associated with the identification of PH7CH as the recurrent parent of the backcross conversion was calculated as 0.99.

6. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Date: 2-28-03 By: Stephen Smith

Stephen Smith

Attorney Docket No. 1328

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant:	Joseph Kevin Gogerty	Date:	Dec. 10, 2002
Serial No.:	09/758,867	Group Art Unit:	1638
Filed:	January 11, 2001	Examiner:	David T. Fox
For:	"INBRED MAIZE LINE PH7CH"		

Assistant Commissioner for Patents
Washington, D.C. 20231

RULE 132 DECLARATION
OF
DR. STEPHEN SMITH

Sir:

I, Stephen Smith, PhD., do hereby declare and say as follows:

1. I am skilled in the art of the field of the invention. I have a Ph.D. in Biochemical Systematics and Taxonomy of Maize and its Wild Relatives from Birmingham University. I have a M.Sc. in the Conservation and Utilization of Plant Genetic Resources from Birmingham University. I have a Bachelor of Science degree in Plant Sciences from London University. Since 1977 I have been engaged in the development, study and application of molecular markers to genetics, measuring genetic diversity and tracking pedigrees. I commenced this work at North Carolina State University as a post-doctoral research fellow. I have continued my engagement in these studies during my employment by Pioneer Hi-Bred from 1980 until the present. These studies have resulted in numerous scientific articles that have appeared in peer reviewed scientific literature.

2. I have read and understood the Office Action in the above case dated October 30, 2002. This declaration is in response to the Examiner's rejection under, 35 U.S.C. § 102(e) as anticipated by or, in the alternative, under 35 U.S.C. § 103(a) as obvious over Garing (U.S. Patent No. 6,034,304).

3. I have conducted an analysis of SSR marker data for inbred PH7CH and the inbred cited as prior art, 90LDC2. Out of a total of 70 SSR loci examined, which allowed a sampling of each chromosome, there are 41 markers that show differences between PH7CH and 90LDC2. This represents a difference for 59% for the markers tested. Of

Appendix G

these 41 markers, 22 were greater than 50 cM in distance, or unlinked on the genetic map.

4. Upon crossing PH7CH to any other maize line and selfing successive filial generations, one would within the realm of what is statistically possible, obtain a progeny inbred maize line that retains genetic contribution from PH7CH. Assuming that (i) the cited prior art is used as the maize line to which PH7CH is crossed, (ii) that the only difference between PH7CH and 90LDC2 are these 41 markers, and (iii) that all markers within a 50 cM distance will segregate together, then the odds of obtaining a PH7CH progeny inbred that is the same as 90LDC2 after one cycle of breeding, is 1 in 2^{22} or 1 in 4,194,304. Statistically it is extremely unlikely that a PH7CH progeny, after one cycle of breeding, would be the same as 90LDC2.

5. Further, the assumptions made above vastly overstate the likelihood of breeding PH7CH from 90LDC2. For example, it is common practice in quantitative genetics to determine the relation of plants by differences in markers. In doing so, one extrapolates that a percentage difference in markers is indicative of a difference in the whole genome. To assume that the only differences between PH7CH and 90LDC2 are for these 41 markers, when 41 markers constitute 59% of the 70 SSR loci examined, is a gross and unrealistic assumption. Further the current maize genetic map only has approximately sixty 50cM units, so by applying this limitation the maximum number of independently segregating loci one could obtain, using the most different maize lines that could ever be found, is sixty. These assumptions result in an over estimate of the odds of breeding PH7CH from 90LDC2.

6. Given the difference in molecular markers between PH7CH and 90LDC2, it is my expert opinion that PH7CH and 90LDC2 are very distinct inventions. It is also my expert opinion that, within the realm of what is statistically possible, any progeny of PH7CH developed through crossing PH7CH with another plant will be distinct from 90LDC2. Given the facts and based on my education and scientific experience, I believe that the invention as claimed is not obvious nor anticipated by Garing (U.S. Patent No. 6,034,304).

7. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

09/758,867

Date: Dec 11th 2002 By: Stephen Smith

Stephen Smith

ASSESSING PROBABILITY OF ANCESTRY USING SIMPLE SEQUENCE REPEAT PROFILES: APPLICATIONS TO MAIZE INBRED LINES AND SOYBEAN VARIETIES

Donald A. Berry,* Jon D. Seltzer,† Chongqing Xie,‡ Deanne L. Wright,‡ Elizabeth S. Jones,‡ Scott Sebastian,‡ J. Stephen C. Smith‡

* Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030

†Medtronic Inc., Minneapolis, MN 55432

‡Pioneer Hi-Bred International, Johnston, IA 50131

SHORT RUNNING HEAD:

Probability of ancestry using SSR

KEY WORDS:

Inbred-alleles, Parentage, Pedigree, SSR, Bayes' Rule

CORRESPONDING AUTHOR:

Donald A. Berry, Ph.D., Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 447, Houston TX 77030-4009.

Phone: (713) 794-4141

Fax: (713) 745-4940

E-mail: dberry@mdanderson.org

ABSTRACT

Determining parentage is a fundamental problem in biology and in applications such as identifying pedigrees. Difficulties inferring parentage derive from extensive inbreeding within the population, whether natural or planned; using an insufficient number of hypervariable loci; and from allele mis-matches caused by mutation or by laboratory errors that generate false exclusions. Many studies of parentage have been limited to comparisons of small numbers of specific parent-progeny triplets. There have been few large-scale surveys of candidates in which there is no prior knowledge of parentage. We present an algorithm that determines the probability of parentage in circumstances where there is no prior knowledge of pedigree and which is robust in the face of missing data and mis-typed data. The focus is parentage of an inbred line having uncertain ancestry. The algorithm is a variation of a previously published hybrid-focused algorithm. We describe the algorithm and demonstrate its performance in determining parentage of 43 inbred varieties of soybean that have been profiled using 236 SSR loci and from seven inbred varieties of maize that were profiled using 70 SSR loci. We include simulations of additional levels of missing and mis-typed data to show the algorithm's utility and flexibility.

The determination of parentage using molecular marker data has been little addressed for situations where there is little or no prior knowledge of parentage, or when large-scale surveys involving numerous candidate parents are required. Consequently, we have recently developed an algorithm and demonstrated its use in determining probability of parentage for hybrids in circumstances where there is no prior knowledge of pedigree and which is robust in the face of missing or mis-typed data (Berry *et al.* 2002). We now present a variation of this algorithm that allows determination of parentage for inbred lines or homozygous varieties.

We describe and evaluate a methodology that quantifies the probability of parentage of homozygous genotypes. Our algorithm takes into account that generations of self-pollination occur after the initial parental cross. The number of generations and the initial parental genotypes are unknown. Each generation of inbreeding reduces the number of heterozygous loci in the progeny by an average of 50%. Thus, each of the inbred progeny individuals resulting from the initial parental cross will have lost approximately half of the parental alleles for loci where the inbred parents were fixed for alternate alleles and which were heterozygous in the F1 generation.

The loss of parental alleles during the inbreeding phase is in contrast to the case of a hybrid progeny. An inbred progeny individual will exhibit a lower level of allelic similarity to either of its inbred parents than a hybrid progeny will to its inbred parents. This loss of some parental alleles during inbreeding might be expected to make an inbred algorithm less robust in the face of missing or mis-typed data compared with the hybrid algorithm that has been previously described (Berry *et al.* 2002). We therefore demonstrate the effectiveness and robustness of the inbred algorithm using examples from two species of cultivated plants. We first tested the

algorithm using varieties of the naturally self-pollinating, inbred crop, soybean [*Glycine max. (L.) Merr.*]. This crop was selected because numerous varieties of soybean with known pedigrees were available to us, many of which are closely related. We also used publicly bred inbreds of maize (*Zea mays L.*) that are of known pedigree. Maize is naturally an outcrossing species but inbred lines are most usually generated for use as parents of commercial hybrids. Inbred lines are generated by making successive generations of self-pollination following the initial bi-parental cross.

MATERIALS AND METHODS

Algorithm: The algorithm is a variation of the hybrid version of Berry *et al.* (2002). Consider an index inbred whose parentage is unknown or in dispute. A database containing possible inbred ancestors is available. The objective is to find the probabilities of closest ancestry for each inbred in the database using genotypic information from a large number of SSRs.

Consider a pair of possible ancestors, inbred *i* and inbred *j*. We calculate the probability that inbreds *i* and *j* are in the index's ancestry, repeating this for all pairs of inbreds in the database. Let $P(i,j|SSRs)$ stand for the posterior probability that *i* and *j* are ancestors of the index given the information from the various SSRs. Let $P(i,j)$ stand for the unconditional (or prior) probability of the same event and let $P(SSRs|i,j)$ be the probability of observing the various SSR results if in fact *i* and *j* are ancestors of the index. Just as in Berry *et al.* (2002), Bayes' rule relates these various probabilities:

$$P(i,j|SSRs) = P(SSRs|i,j)*P(i,j) / \Sigma[P(SSRs|u,v)*P(u,v)],$$

where the sum in the denominator is over all pairs of inbreds in the database, indexed by u and v. We need to calculate $P(SSRs|i,j)$ for each i and j. We will make the "no-prior-information" assumption that $P(i,j)$ is the same for all pairs (i,j). Then $P(u,v)$ is a common multiple in the denominator that cancels with $P(i,j)$ in the numerator:

$$P(i,j|SSRs) = P(SSRs|i,j) / \Sigma P(SSRs|u,v).$$

The problem is to calculate a typical $P(SSRs|i,j)$, the probability of observing the index's SSRs assuming inbreds i and j are both ancestors. The nature of breeding before the self-pollination process is unknown. Since the creation of an inbred proceeds by multiple generations of self-pollination on a hybrid, we label the (unknown) hybrid used to create the (known) index inbred as the intermediate hybrid. When the intermediate hybrid is an immediate descendent of i and j, it receives one of inbred i's alleles and one of inbred j's alleles. When the intermediate hybrid is a second generation descendent of i and j, it receives one allele from each with probability 0.5. And so on. Since degree of ancestry (if any) is unknown, we label the actual probability of passing on one of these alleles to the intermediate hybrid to be p. As in Berry *et al.* (2002) we consider $p = 0.50$ and $p = 0.99$ and here we also consider the intermediate value $p = 0.75$.

When inbreds i and j are ancestors then there are four possibilities: (1) the alleles of both i and j were passed to the intermediate hybrid, (2) i came through but not j, (3) j came through but not i, and (4) neither came through. Assuming independence, these have respective probabilities p^2,

$p(1-p)$, $p(1-p)$, $(1-p)^2$. An allele in the intermediate hybrid's genotype that did not arise from either inbred *i* or inbred *j* is assumed to be selected with probability $1/n$, where n is the total number of alleles at the SSR in question. So far the steps we have described are identical to those for identifying the ancestors of a hybrid described by Berry *et al.* (2002) and, in fact, if the index is heterozygous at an SSR then calculations proceed just as for hybrids. Calculations are substantially different when the index inbred is homozygous, say genotype *aa*. Cases that must be considered are shown in Table 1, where *x* is any allele different from *a* (but not missing). All alleles other than *a* can be grouped because only *a* appears in the index's genotype. For example, *xx* might be *bc* or *bd* or *bb*.

P(SSR|i,j) is the probability of observing the index assuming inbreds *i* and *j* are ancestors. The calculations for SSRs 1 to 6 are shown in Table 2, where the four terms in each case are in order of (1), (2), (3), (4) defined in the previous paragraph. Missing alleles are not considered in the examples above. The number of possibilities is large. Here we consider only the case in which inbred *i* is *aa* and both alleles of inbred *j* are missing. Then

$$P(SSR|i,j) = p^2(1/2+1/2*1/n) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/n) + (1-p)^2(1/n)$$

Another possibility not considered above is that more than two alleles can be observed for an SSR marker run on individual DNA sample. This can be due to SSR locus duplication, homology due to alloploidy, more than one individual plant being sampled for DNA extraction or cross-contamination. In this case we consider all possible pairings of the observed alleles and calculate using a multiple imputation procedure (Little and Rubin, 1987).

To find the overall $P(SSRs|i,j)$, multiply the individual $P(SSR|i,j)$ over the various SSRs. To determine the probability that any particular inbred, say inbred i, is the closest ancestor of the index, sum $P(SSR|i,v)$ over all inbreds v with $v \neq i$. Call this $P(i|SSRs)$. The maximum of $P(i|SSRs)$ for any inbred i is 1. But since there is one closest ancestor on each side of the family, the sum of $P(i|SSRs)$ over all inbreds is 2.

SSR data: Soybean DNA was extracted from 490 varieties, all of which were bred in, and are adapted to, the United States. Plant material for DNA extraction was sampled from six plants of each variety. Most of the varieties are proprietary products of Pioneer Hi-Bred International. Several (non-patented) commercial varieties from other breeding companies and some important publicly bred varieties were also included. Procedures for obtaining SSR data from soybean were identical to those described for maize by Berry *et al.* (2002) apart from the following modifications: PCR products with different size ranges and labeled with different fluorochromes were pooled and diluted 1:9 with capillary electrophoresis buffer (Applied Biosystems) then 1:4 with dH20. 1.5ul of pooled DNA were added to 10ul formamide containing the molecular weight size standard 400HD ROX (Applied Biosystems, ROX = 6-carboxy-X-rhodamine). Fragment separation was performed using capillary electrophoresis on an ABI3700 platform (Applied Biosystems), with an injection time of 10 sec at 10,000 V and a run time of 4,000 sec at 7,500 V. Forty-three soybean varieties that had both of their parent varieties also included in the dataset were assigned as index varieties. One to two and occasionally three grandparent varieties of several of the index varieties were also included in the dataset. These varieties collectively

represent a broad array of diversity of soybean germplasm that is currently grown in the United States.

Two hundred and thirty-six publicly available soybean SSR markers (http://soybase.agron.iastate.edu/) were used to demonstrate and evaluate the algorithm. These SSR markers were selected following initial screens on a subset of 24 soybean varieties in which they were tested for amplification and the ability to detect polymorphism. The 236 markers gave good genome coverage and collectively mapped across each of the chromosomal linkage groups of soybean.

All allele scores were made without knowing the identities of the soybean genotypes.

Maize SSR data using 70 loci were previously reported by Senior *et al.* (1998) and were obtained directly from the first author. This publication (Senior *et al.* 1998) cites an array of 94 historically important publicly bred lines that have well known and well established pedigrees. This array of public inbreds includes seven inbreds (A632; A634, Mo17, Pa91, Va35, Va99 and W64A) that each have SSR profiles for their parental lines included in the same dataset. Three of these inbreds were developed from a breeding cross of two unrelated parents. These are: Mo17 which was bred from the cross of C.I. 187-2 x C103; Va99, which was bred from the cross Oh07B x Pa91; and W64A which was bred from the cross of WF9 x C.I. 187-2. Other inbred progeny had more complex pedigrees. One inbred (Va35) was bred from the cross C103 x T8 following an additional cross of T8 as the recurrent parent. Two inbreds (A632 and A634) were bred from the cross Mt42 x B14 following additional crosses of B14 as the recurrent parent.

Pa91 was bred from a complex cross involving four inbreds (WF9 x Oh40B) and (38-11 x L317). These seven progeny inbreds therefore provided an index set of maize inbreds for evaluation of the inbred algorithm.

RESULTS

Data quality: The soybean SSR data that were used to evaluate the algorithm had a mean of 5.5% (range 0-19% loci) missing data per variety. For parent-progeny triplets, there was a mean of 1.1% loci (range 0-5%) where a progeny profile was scored for an allele that was not represented by either of the seed sources that represented the parents. The maize SSR data had a mean of 0.7% missing data (only three genotypes had missing data; these were at elevated levels of 5%, 9%, and 36%). A mean of 6.4% parent/progeny triplets (range 4-7%) had SSR progeny profiles that did not share an allele with either of the seed sources that were available to represent the original parental genotypes.

Probability of ancestry applied to soybean data: Figures 1 and 2 present the probabilities of closest ancestry of the top ranking varieties for each of 43 soybean varieties using data from 236 marker loci at $p = 0.50$ (Fig 1) and at $p = 0.99$ (Fig 2).

When the algorithm was used at $p = 0.5$ with data from all 236 loci (Fig 1), then 24/43 (56%) of index varieties had both parents correctly identified in the top two ranked positions, 12/43 (28%) had one parent correctly placed in one of the top two positions, and 7/43 (16%) had none of the actual parents assigned into the top two ranked positions. Thus, when $p = 0.5$ was used, 60/86

(70%) of actual parental varieties were correctly ranked in the top two positions and 26/86 (30%) were incorrectly placed in lower positions.

When the algorithm was used at $p = 0.75$ with data from all 236 loci (data not shown), 28/43 (65%) of index varieties had both parents correctly identified in the top two ranked positions, 11/43 (26%) had one parent correctly placed in one of the top two positions, and 4/43 (9%) had none of the actual parents assigned into the top two ranked positions. Therefore, when $p = 0.75$ was used, 67/86 (78%) of the actual parental varieties were correctly ranked in the top two positions and 19/86 (22%) were incorrectly placed in lower positions.

When the algorithm was used at $p = 0.99$ with data from all 236 loci (Fig 2), then 33/43 (77%) of actual parental varieties were correctly ranked in the top two positions and 10/86 (23%) had one parent correctly placed; all index varieties had at least one parent ranked in the top two positions when the algorithm was used at $p = 0.99$. With p used at 0.99 then 76/86 (88%) of actual parental varieties were correctly assigned; 10/86 (12%) were incorrectly assigned.

Table 3 presents the rankings, probabilities, and pedigrees of varieties that were incorrectly assigned above a true parent. The largest pedigree class (41% of cases where a non-parent ranked above a true parent) of non-parents ranking higher than parents was for varieties that are derivatives of the parent that was misplaced at a lower ranking. The equal second largest classes (each representing 14% of the cases) were for varieties that were (a) full sibs of the true but misplaced parent and (b) full sibs of a grandparent of the variety for which the pedigree was being tested. Other categories (percent of cases in parentheses) were: multiple backcross versions

of the misplaced parent (7%), a derivative of the variety or which the pedigree was being tested (7%), a half-sib of the true but lower ranked parent (7%), a full sib of the variety for which the pedigree was being tested (3%), and a half-sib of the variety for which the pedigree was being tested (3%). Insufficiently detailed pedigree information is available to categorize the variety (3% of cases) that ranked above the true parent

Robustness: The quality of soybean SSR data as received from the laboratory, in terms of missing data and apparently non-Mendelian parent-progeny triplets, have already been presented. Taking these data as an initial starting point, additional levels of missing and mis-typed data were created by simulations and used to explore robustness of the algorithm.

SSR data for five index soybean varieties were used to determine the robustness of the algorithm. Subsets of data were created that included parameters of reduced numbers of loci, additional levels of missing data; additional levels of mis-typed data, and various combinations of these parameters. Simulated levels of missing and mis-typed data were created with a first pass creating missing data, followed by a second pass creating mis-typed data. Therefore, for example, the maximum level of cumulative error from simulated missing and mis-typed data was from 36 to 40%. Five varieties were chosen to represent a range of diversity in respect of both pedigree and SSR profiles. Four varieties had no parents or grandparents in common and one pair of varieties was related by a common parent. All varieties had parents ranked in the top two positions when the algorithm was run at $p = 0.75$ and $p = 0.99$. This selection of varieties therefore provides a means to establish lower boundaries for both the quantity and quality of SSR data that are required to avoid aberrant results.

Table 4 presents the probability of ancestry of the top five ranked varieties for each of five selected soybean index varieties (93B11, A7986, P9443, S38T8 and Young) when the algorithm is run using different numbers of SSR marker loci (50, 100, 150 and 236) at each of two levels of p (0.5 and 0.99). Using $p = 0.5$, the lowest percentage of parents (60%) that were correctly ranked into the top two positions corresponded to using only 50 SSR. Increasing the number of loci to 100 or 150 or 236 increased the ability to identify the actual parents to about 90%. When p was used at a level of 0.99 all parents were correctly ranked into the top two positions for each of the five varieties when data from as few as 50 SSR loci were used.

Table 5 summarizes other aspects of robustness. Namely, we simulated additional levels of missing, mis-typed and missing plus mis-typed data, beyond those that were inherent in the data as provided by the laboratory. When p was used at a level of 0.5, robustness was generally maintained up to an additional level of 20% simulated missing data, so long as data from 100 or more loci were used. Similarly, robustness was maintained for up to 20% additional mis-typed data so long as data from 100 or more loci were used. Likewise, robustness was maintained with up to 18 to 20% additional levels of data error including both missing and mis-typed data, so long as data from 150 or more loci were used. Using data for all 236 loci provided a higher level of robustness, but even then robustness collapsed when 36 to 40% cumulative additional error from missing and mistyped data were simulated into the analysis. The overall level of correct assignation of parent varieties was higher when p was used at a level of 0.99. All parents then were correctly identified, even when data from only 50 loci were used up to an additional level of 10% missing data. When data from 100 or more loci were used then all parents were correctly

identified with up to 20% additional missing data. Robustness started to decline when the algorithm was applied with 10% additional mis-typed data when data from 150 or fewer SSR loci were used. However, robustness was maintained for up to 20% additional mis-typed data when data from 236 SSR loci were used. When additional levels of both incorrect data were applied then robustness was maintained at levels of up to 10% missing plus 10% mis-typed data so long as data from at least 150 SSR loci were used. Robustness was compromised when additional simulations of 20% missing plus 20% mis-typed data were applied even when data from all 236 SSR loci were used.

We then investigated the relationships of varieties to the index genotype whose pedigree was under examination by rerunning the analysis after both parents of the index genotype had been removed from the analysis. Fifteen varieties that had two or more of their grandparents profiled in the dataset were used for this examination. After removing parents, direct pedigreed derivatives of the index genotype ranked first for P9583, in the first three places for A2943 and in the first six places for P9561. Once all parents and derivatives of the index genotype had been removed from the analysis then the following results were obtained. Predominant classes of varieties ranking in the top five positions were (percent of cases in parentheses): derivatives of the grandparent of the index variety (32%), grandparents of the index variety (16%), derivatives of the parents of the index variety (16%), and half-sibs of the index variety (13%). Grandparents ranked among the first four positions for 10 varieties and were in the first place for five varieties. Great-grandparents ranked within the first seven places for three varieties, and a great-great-grandparent ranked in eighth place for one variety. Other varieties that ranked in the first place were usually closely related to the variety whose pedigree was under examination; full-sibs and

half-sibs were the predominant classes of relatives other than grandparents in the first ranking position after parents and direct derivatives of the variety under examination had been removed.

Probability of ancestry applied to corn data: The seven index inbreds of maize were selected because they represented all of the inbred lines published upon by Senior *et al.* (1998) that had all of their inbred parents also included in the SSR dataset. All of the inbred lines published by Senior *et al.* (1998) have well known and well established pedigrees that are fully provided by those authors.

Table 6 presents probabilities of ancestry for the top five ranked inbreds for each of the seven index inbred lines at two levels of p (0.5 and 0.99). For the three progeny that were bred from single crosses without any subsequent use of one of the parents to make a recurrent cross prior to inbreeding (Mo17, Va99, and W64A) then use of the algorithm at either $p = 0.5$ or at $p = 0.99$ resulted in the parental inbreds being ranked in first and second positions. Use of the algorithm at $p = 0.99$ provided greater discrimination for probabilities of ancestry that were assigned to actual parents compared to highest ranking non-parents. This was most noticeable for the case of inbred Va99 which had a relatively low value when used at $p = 0.5$ for parent 2 (0.5221) compared to parent 1 (0.9999) or to the third ranked inbred (and non-parent), Va22 (0.4252). In contrast, when the program was run at $p = 0.99$ then parent1 and parent2 for Va99 had probabilities of 1 and 0.9855, respectively, with the probability of the third ranked inbred being 0.0131.

For each of the three progeny inbreds that originated from breeding schemes that involved one or more additional crosses of one of their parents, using the algorithm at $p = 0.5$ resulted in

placement of the respective recurrent parent with the highest probability of ancestry. Raising the level of p to 0.99 resulted in both parents (B14 = recurrent parent and MT42 the non-recurrent parent) of the index inbred A632 being ranked in the top two places. Using this level of p also caused a higher ranking (third position) for the non-recurrent parent (MT42) of index inbred A634. Use of p at 0.99 did not cause the non-recurrent parent (C103) of index inbred (Va35) to rank into the top five places.

For the index inbred (Pa91) that was bred from a more complex cross involving four inbred lines, the use of p at 0.5 or at 0.99 resulted in the two parents (WF9 and Oh40B) being ranked in second and third places; highest ranked was inbred Va99 (Va99 is derived from the index inbred Pa91). Neither of the two remaining parents of Pa91 ranked in the top five places.

DISCUSSION

The current widely used North American soybean varieties are founded upon a relatively narrow genetic base of diversity. Gizlice *et al.* (1994) document that the U. S. soybean germplasm base is founded upon 20 plant introductions and that subsequent breeding has made repeated use of related parents. Molecular marker comparisons of elite U. S. soybean varieties compared to a sample of exotic varieties reinforce the conclusion that there is a relative paucity of genetic variation in U. S. soybeans. Narvel *et al.* (2000) have shown that the number of alleles detected among the exotics was 30% greater than among U. S. varieties. Thompson and Nelson (1998) report that very little exotic germplasm has been incorporated into the existing U. S. soybean germplasm base. Examining all pairs of pedigree relationships among the 490 soybean varieties

employed in this study showed that approximately 50% of pairwise relationships are related at the level of half-sib or closer; approximately 10% of pairs are related at the level of full-sib or closer. This set of soybean varieties therefore provides the basis for an extremely rigorous evaluation of the ability of SSR data to distinguish between varieties and of this algorithm to identify pedigrees. Pedigree breeding, including the use of related parents, is also commonly applied in the breeding of maize inbred lines. The set of maize inbreds used here thus also provides a meaningful evaluation of the marker data to discriminate among inbred lines and of the joint ability of the algorithm and of the marker data to allow a determination of inbred pedigrees.

Use of the algorithm at $p = 0.99$ rather than at a lower level improved performance in terms of the percentage of correct assignations of parents and provided a greater statistical differential for probabilities for parents in comparison to the highest ranking non-parents. Use of the algorithm at $p = 0.99$ is more appropriate when it is known that the actual parents of the variety under examination are included among the set of index varieties. If it is not known that the parents are included in the index set then use of the algorithm at $p = 0.5$ is more justified (Berry *et al.* 2002). For the soybean varieties, when p was used at 0.99, then 77% of all varieties that were queried for their parents had both parents correctly identified. Eight-eight percent of soybean parents were correctly identified across 43 index varieties that were queried for their parents. All varieties (with the possible exception of one variety where detailed pedigree information was not available) that ranked above true parents were related either to the mis-ranked parent or to the variety that was being queried for its pedigree. Our previous report of the use of an algorithm to determine hybrid pedigrees (Berry *et al.* 2002) showed a higher level of correct parental

determinations at $p = 0.99$. Many of these soybean varieties have a high degree of pedigree relatedness. However, many of the maize inbred lines that were used in the previously reported study (Berry *et al.* 2002) were also highly related. It is, however, likely to be inherently more challenging to correctly identify parents following cycles of inbreeding because half of the alleles that are segregating in the first generation following the initial breeding cross will be subsequently lost as recurring cycles of self-fertilization occur. Thus, many of the alleles that are present in a hybrid, and which can therefore contribute to the identification of its pedigree, do not remain present in an inbred homozygous progeny.

We examined the pedigrees of soybean index varieties when both parents of the index had been removed from the set of candidate varieties. Direct pedigree descendents with the index variety as one parent then usually ranked higher than other varieties, including varieties that were grandparents or sister varieties of the index variety. When all parents and direct derivatives of the index variety were excluded from the analysis then the predominant classes of varieties ranking in the top five positions were derivatives of the grandparent of the index variety (32%), grandparents of the index variety (16%), derivatives of the parents of the index variety (16%), and half-sibs of the index variety (13%). The SSR data that were available to us did not allow a thorough or very precise assessment of how varieties with different degrees of relatedness would rank as members of the pedigree in the event that the true parents were not present in the database. Nonetheless, when parents were excluded from the analysis then varieties that were very closely related to the index variety ranked highest. Direct descendents dependent for their pedigree upon the index variety, if present, tended to rise above varieties included within other classes of pedigree relationship to the index variety. When varieties directly descended by

pedigree from the index variety were also excluded then a grandparent ranked into first position for 33% of the varieties that were examined. Direct pedigree derivatives of one or more of the parents of the index variety had an equal level of occurrence when parents and derivatives of the index variety were excluded. Further investigations of the identification of grandparents will require a dataset including all grandparents of each index variety and will also require a revised algorithm to take account of pedigree contributions from four varieties as opposed to pairs of varieties which forms the basis of the current inbred algorithm.

For the maize inbred line pedigrees, use of the algorithm either at $p = 0.5$ or at $p = 0.99$ resulted in the correct identification of both parents in all cases where the breeding scheme was an initial cross of two parental lines followed by subsequent cycles of inbreeding (i.e. for the inbreds Mo17, Va99 and W64A). The relatively high level of robustness for results with maize inbreds at $p = 0.5$, in contrast to the results obtained from analyzing soybean data (where 56% of varieties had both parents correctly identified when p = 0.5 was used) could be accounted for by the smaller sample size of maize inbreds and by the lower degree of mean pedigree relatedness amongst this selection of inbred lines in comparison to the soybean varieties. Thus while several inbred lines in this set are closely related, there remain many inbreds that have little or no pedigree relationship (Senior *et al.*, 1998).

The inbred algorithm correctly identified both parents of the three maize index inbreds that had been bred from bi-parental crosses that involved equal contributions (by pedigree) from both parents. For the three bi-parental crosses that involved subsequent additional crosses of the recurrent parent (and thus significantly biased contributions by pedigree to the index variety

from the recurrent parent) then use of the algorithm correctly identified each of the recurrent parents. The algorithm was unable to identify the non-recurrent parent in most cases, but this result would be expected because one backcross reduces the expected pedigree contribution of the non-recurrent inbred to 25%. More generations of backcrossing using the recurrent parent then further reduce the expected pedigree contribution of the non-recurrent parent by half at each generation (successively to 12.5%, 6.25%, 3.125%) with the pedigree contribution of the recurrent parent rising accordingly. Since several inbred lines of maize are related by pedigree then it is not surprising that the level of pedigree or SSR similarity of a non-recurrent parent to the index progeny can fall below other inbred lines that are related to the index variety. The algorithm was not able to preferentially identify parents of the inbred line Pa91, which was bred from a complex breeding scheme involving four parents with equal contributions by pedigree. A more suitable algorithm is needed to take account of four way crosses. However, such a need is primarily academic because most breeding crosses in commercial maize breeding, and indeed for most crops, are bi-parental.

These soybean data had a mean of 5.5% missing data per variety and a mean of 1.1% loci where a progeny was scored with an allele that was not also scored in either or both parents. Such apparent non-Mendelian or exclusionary profiles can be due to pollen contamination during inbreeding, cross contamination in the field or laboratory, scoring errors in the laboratory (e.g. scoring +A, predominant stuttering, spectral pull-up, secondary binding sites or polymer spikes), or incorrect pedigrees. Another source of apparent exclusion is through the use of a seed source as a parent that is still heterogeneous due to inbreeding being incomplete. Cycles of inbreeding then continue so that when those seed sources are used in the future as sources for SSR profiling

to represent the parental genotype they will have lost alleles due to inbreeding that have already been passed on to a progeny. Alternately, residual heterozygosity within seed sources can result in low frequencies of heterozygotes or off-type segregants which may, by chance, be sampled in the progeny, but not sampled in the parent. In this study we sampled six plants to represent the variety which may be insufficient to capture alleles existing at low frequencies within the seed source. And even if the allele was sampled, it may not have been detected following PCR amplification due to predominance of the most frequent allele and allelic competition effects. Hall (2002) has also reported the occurrence of apparent non-parental SSR alleles. Mutation can also affect SSR profiles. Vigouroux *et al.* (2002) have estimated mutation rates of 7.7×10^{-4} per generation for dinucleotide SSRs and an upper 95% confidence limit of 5.1×10^{-5} for SSRs with longer repeat units. A level of error or discrepancy in expected SSR profiles are thus inevitable for some, if not all crop plants. We therefore evaluated the robustness of the algorithm and dataset by rerunning the algorithm using datasets that were simulated to have up to 20% additional levels of missing plus 20% mis-typed data beyond the level that was received from the laboratory. The algorithm maintained its initial level of robustness with up to an additional level of 10% both missing and mis-typed data, provided data from at least 100 SSR loci were used. Fewer loci (60) were capable of retaining this degree of robustness in the evaluation of the hybrid pedigree algorithm using maize hybrids (Berry *et al.* 2002). The loss of parental alleles that occurs during the inbreeding process, in contrast to their retention in a hybrid progeny compared to its parents, probably underlies the need to use data from a greater number of loci to maintain robustness for the inbred algorithm as compared to the hybrid algorithm.

It was anticipated that determination of pedigrees following cycles of inbreeding might be more challenging to accomplish than to determine pedigrees of hybrids where the total nuclear genetic contributions of both parents are preserved. Nonetheless, these results show that the algorithm can be used effectively to identifying parents of inbred genotypes. Nearly 90% of soybean parents were identified. This is a set of genotypes which, due to the relatively narrow founder base and subsequent cycles of development through the use of related crosses, provides an extremely rigorous test of the algorithm and of the discriminatory power of the marker data. Supplementary data also show the capability of the algorithm to identify parents of maize inbreds that have been developed in a pedigree system using two parents. Use of this algorithm with currently available codominantly expressed molecular marker data has also been shown to have practical feasibility because of the high degree of robustness that is evident and which extends well beyond the realm of aberrant or unexpected marker data that is encountered. These types of error or unexpected marker data can include laboratory error, sampling effects or the use of different seed sources for the actual parental source compared to a more inbred source that becomes available later to represent the parental genotype. This algorithm has application in a number of fields, including conservation biology, population genetics, and to assist in the protection of intellectual property rights.

LITERATURE CITED

Berry, D. A., J. D. Seltzer, C. Xie, D. L. Wright, and J. S. C. Smith, 2002 Assessing probability of ancestry using simple sequence repeat profiles: applications to maize hybrids and inbreds. Genetics 161: 813-824.

Gizlice, Z., T. E. Carter Jr., and J. W. Burton, 1994 Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34: 1143-1151.

Hall, M.A., 2002 Inbred corn plant 01HF13 and seeds thereof. Patent No, US 6,353,161 B1. U.S. Patent Office, Washington DC.

Little, R. J. A. and D. B. Rubin, 1987 *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York.

Narvel, J. M., W. R. Fehr, W-C Chu, D. Grant, and R. C. Shoemaker, 2000 Simple sequence repeat diversity among soybean plant introductions and elite genotypes. Crop Sci. 40: 1452-1458.

Senior, M. L., J. P. Murphy, M. M. Goodman, and C. W. Stuber, 1998 Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Sci. 38: 1088-1098.

Thompson, J. A. and R. L. Nelson, 1998 Utilisation of diverse germplasm for soybean yield improvement. Crop Sci. 38: 1362-1368.

Vigouroux, Y., J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. D. Beavis, J. S. C. Smith, and J. Doebley, 2002. Rate and pattern of mutation at microsatellite loci in maize. Mol. Biol. Evol. 19: 1251-1260.

Table 1. Calculations of ancestry for homozygous index inbreds: Cases that must be considered for example of genotype *aa*.

SSR	Index	Inbred *i*	Inbred *j*
1	*aa*	*aa*	*Aa*
2	*aa*	*aa*	*Ax*
3	*aa*	*aa*	*Xx*
4	*aa*	*ax*	*Ax*
5	*aa*	*ax*	*Xx*
6	*aa*	*xx*	*Xx*

x is any allele different from *a*, but not missing

Table 2. Probability of observing the index [$P(SSR|i,j)$] assuming inbreds i and j are ancestors: Calculations for SSRs 1 to 6.

SSR	$P(SSR\|i,j)$
1	$p^2(4/4) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/2+1/n*1/2) + (1-p)^2(1/n)$
2	$p^2(3/4) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/2*1/2+1/n*1/2) + (1-p)^2(1/n)$
3	$p^2(2/4) + p(1-p)(1/2+1/n*1/2) + p(1-p)(1/n*1/2) + (1-p)^2(1/n)$
4	$p^2(2/4) + p(1-p)(1/2*1/2+1/n*1/2) + p(1-p)(1/2*1/2 + 1/n*1/2) + (1-p)^2(1/n)$
5	$p^2(1/4) + p(1-p)(1/2*1/2+1/n*1/2) + p(1-p)(1/n*1/2) + (1-p)^2(1/n)$
6	$p^2(0/4) + p(1-p)(1/n*1/2) + p(1-p)(1/n*1/2) + (1-p)^2(1/n)$

The four terms in each case are in order of the four possibilities when inbreds i and j are ancestors: (1) the alleles of both i and j were passed to the intermediate hybrid, (2) i came through but not j, (3) j came through but not i, and (4) neither came through. Missing alleles are not considered.

Table 3. Probabilities of ancestry and pedigree relationships for soybean varieties where both parents did not rank above non-parents.

Case no.	Index variety	Rank	Possible ancestor	Probability
1	*95B97*	1	Parent 2	1
		2	Full sib of parent 1	0.5822
		3	Parent 1	0.4124
2	*A2943*	1	Parent 1	0.9977
		2	Multiple backcross of parent 2	0.7999
		3	Parent 2	0.1999
3	*A4595*	1	Parent 1	1
		2	Derivative of parent 2	0.9956
		3	Multiple backcross of parent 2	0.0034
		4	Derivative of Parent 2	0.0006
		5	Half sib of A4595	0.0004
		6	Parent 2	0.0001
4	*Hark*	1	Parent 1	1
		2	Derivative of parent 2	1
		3	Derivative of parent 2	2.1E-09
		4	Derivative of parent 2	1.4E-09
		5	Derivative of Hark	3.1E-10
		6	Derivative of parent 2	1.1E-13
		7	unknown	3.8E-15
		8	Derivative of parent 2	4.6E-17
		9	Derivative of parent 2	4.7E-21
		10	Parent 2	2.7E-21
5	*Kent*	1	Parent 2	1
		2	Derivative of parent 1	0.9990
		3	Derivative of parent 1	0.0011
		4	Parent 1	3.0E-04
6	*P9583*	1	Parent 1	1
		2	Full sib of P9583	0.8801
		3	Parent 2	0.1199
7	*P9641*	1	Parent 2	1
		2	Derivative of P9641	1
		3	Parent 1	3.7E-06
8	*S30J2*	1	Parent 1	1
		2	Derivative of parent 2	0.9321

		3	Parent 2	0.0679
9	*YB30K01*	1	Parent 2	1
		2	Half sib of parent 1	1
		3	Full sib of parent 2	7.9E-09
		4	Half sib of parent 2	3.3E-09
		5	Full sib of grandparent	1.2E-10
		6	Derivative of parent 1	3.0E-11
		7	Full sib of parent 2	2.0E-11
		8	Full sib of grandparent	8.7E-12
		9	Parent 1	1.1E-12
10	*YB41Q01*	1	Parent 2	1
		2	Full sib of parent 1	1
		3	Full sib of grandparent	7.3E-05
		4	Full sib of grandparent	4.1E-09
		5	Parent 1	9.1E-10

Results for 33 (77%) varieties where both parents were ranked first and second are not included in this table (see Figures 1 and 2).

Table 4. Probability of ancestry for five individual soybean varieties using SSR data obtained from different numbers of loci (50, 100, 150, 236).

	L50		L100		L150		L236	
Inbred	Possible ancestor	Prob	Possible ancestor	Prob	Possible ancestor	Prob	Possible ancestor	Prob
P=0.5								
93B11	*XB31C*	0.9461	*XB31C*	1	*XB31C*	1	*XB31C*	1
	A3415	0.8006	*A3415*	0.9362	*A3415*	0.9146	*A3415*	0.9954
	XB38A01	0.0256	*WILLIAMS*	0.0429	*WILLIAMS*	0.0809	*WILLIAMS*	0.0046
	P9271	0.0251	*A3242*	0.0155	*YB30L01*	0.0034	*A3242*	0
	YB30L01	0.0232	*YB30L01*	0.0015	*A3242*	0.0006	*DOUGLAS*	0
A7986	*COOK*	0.7748	*BRAXTON*	0.9725	*BRAXTON*	1	*BRAXTON*	1
	XB63D00	0.2841	*YOUNG*	0.5302	*YOUNG*	0.8910	*YOUNG*	0.9929
	S6262	0.1826	*COOK*	0.3872	*P9641*	0.0404	*XB63D00*	0.0071
	YOUNG	0.1755	*XB63D00*	0.0496	*XB63D00*	0.0254	*96B32*	0
	BRAXTON	0.1065	*P9641*	0.0328	*COOK*	0.0245	*P9641*	0
P9443	*DOUGLAS*	0.8086	*A3415*	0.5557	*FAYETTE*	0.8760	*FAYETTE*	0.9885
	A3415	0.7629	*FAYETTE*	0.4957	*A3415*	0.7034	*DOUGLAS*	0.8847
	WILLIAMS	0.0887	*DOUGLAS*	0.4855	*CX399*	0.1671	*A3415*	0.0846
	YALE	0.0501	*CX260C*	0.2032	*CX260C*	0.1273	*WILLIAMS*	0.0348
	P9394	0.0411	*WILLIAMS*	0.1608	*WILLIAMS*	0.0948	*CX399*	0.0062
S38T8	*S3535*	0.8711	*S3535*	0.9993	*S3535*	1	*S3535*	1
	S4644	0.4543	*S4644*	0.9988	*S4644*	1	*S4644*	1
	YB44R01	0.2762	*YB40M01*	0.0012	*YB37Y00*	0	*A4268*	0
	YB40M01	0.1087	*YB44R01*	0.0004	*93B65*	0	*YB44R01*	0
	YB44Q01	0.0325	*YB37Y00*	0.0001	*A4268*	0	*YB37Y00*	0
YOUNG	*DAVIS*	0.6589	*DAVIS*	0.6551	*DAVIS*	0.6324	*DAVIS*	0.9752
	XB63D00	0.4942	*ESSEX*	0.5979	*P9641*	0.5524	*P9641*	0.5397
	96B32	0.3122	*P9641*	0.3409	*COOK*	0.3231	*ESSEX*	0.3273

	COOK	0.0707	*COOK*	0.1692	*ESSEX*	0.2817	*96B32*	0.1299
	OGDEN	0.0606	*96B32*	0.1315	*96B32*	0.1933	*COOK*	0.0235
p=0.99								
93B11	*XB31C*	1	*XB31C*	1	*XB31C*	1	*XB31C*	1
	A3415	0.9999	*A3415*	0.9999	*A3415*	1	*A3415*	1
	A3242	0.0001	*A3242*	0.0001	*P9443*	0	*WILLIAMS*	0
	P9443	0	*P9443*	0	*A3242*	0	*A3242*	0
	WILLIAMS	0	*WILLIAMS*	0	*WILLIAMS*	0	*FAYETTE*	0
A7986	*BRAXTON*	1	*BRAXTON*	1	*BRAXTON*	1	*BRAXTON*	1
	YOUNG	0.9903	*YOUNG*	0.9903	*YOUNG*	0.9987	*YOUNG*	1
	P9641	0.0092	*P9641*	0.0092	*96B32*	0.0012	*XB63D00*	0
	96B32	0.0005	*96B32*	0.0005	*P9641*	0.0002	*96B32*	0
	DAVIS	0	*DAVIS*	0	*DAVIS*	0	*P9641*	0
P9443	*DOUGLAS*	0.9998	*DOUGLAS*	0.9999	*FAYETTE*	0.9995	*DOUGLAS*	1
	FAYETTE	0.7010	*FAYETTE*	0.7011	*DOUGLAS*	0.9993	*FAYETTE*	1
	CX260C	0.2345	*CX260C*	0.2345	*CX399*	0.0006	*CX260C*	0
	A3415	0.0644	*A3415*	0.0643	*A3415*	0.0005	*CX399*	0
	S3941	0.0001	*AP3330*	0.0001	*P9394*	0.0001	*A3415*	0
S38T8	*S3535*	1	*S3535*	1	*S3535*	1	*S3535*	1
	S4644	1	*S4644*	1	*S4644*	1	*S4644*	1
	YB40M01	0	*YB40M01*	0	*93B67*	0	*A4268*	0
	YB44R01	0	*A5979*	0	*ST3780*	0	*YB54J00*	0
	93B67	0	*YB44R01*	0	*YB37Y00*	0	*YB44R01*	0
YOUNG	*DAVIS*	1	*DAVIS*	1	*DAVIS*	1	*DAVIS*	1
	ESSEX	1	*ESSEX*	1	*ESSEX*	1	*ESSEX*	1
	P9641	0	*P9641*	0	*COOK*	0	*S4240*	0

#9 attach

Assessing Probability of Ancestry Using Simple Sequence Repeat Profiles: Applications to Maize Hybrids and Inbreds

Donald A. Berry,*[,1] Jon D. Seltzer,[†] Chongqing Xie,[‡] Deanne L. Wright[‡] and J. Stephen C. Smith[‡]

*The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, [†]Third Wave Technologies, Inc., Madison, Wisconsin 53719 and [‡]Pioneer Hi-Bred International, Inc., Johnston, Iowa 50131

Manuscript received July 24, 2001
Accepted for publication March 11, 2002

ABSTRACT

Determination of parentage is fundamental to the study of biology and to applications such as the identification of pedigrees. Limitations to studies of parentage have stemmed from the use of an insufficient number of hypervariable loci and mismatches of alleles that can be caused by mutation or by laboratory error and that can generate false exclusions. Furthermore, most studies of parentage have been limited to comparisons of small numbers of specific parent-progeny triplets thereby precluding large-scale surveys of candidates where there may be no prior knowledge of parentage. We present an algorithm that can determine probability of parentage in circumstances where there is no prior knowledge of pedigree and that is robust in the face of missing data or mistyped data. We present data from 54 maize hybrids and 586 maize inbreds that were profiled using 195 SSR loci including simulations of additional levels of missing and mistyped data to demonstrate the utility and flexibility of this algorithm.

DETERMINATION of parentage is fundamental to the study of reproductive and behavioral biology. The increasing availability of highly discriminant genetic markers for many diverse species provides the potential to uniquely characterize individuals at numerous loci and to unambiguously resolve parentage where genealogical relationships are unknown, in error, or in dispute.

Identification of parent-progeny relationships in wild populations of animals and plants provides insights into the success of various reproductive strategies (Ellstrand 1984; Smouse and Meagher 1994; Alderson *et al.* 1999) and has allowed for the implementation of management programs to conserve genetic diversity (Miller 1975; Rannala and Mountain 1997). The association of pedigree with physical appearance or performance in domesticated animals and plants allows parents that have contributed favorable alleles for desirable traits through selective breeding programs to be identified (Bowers and Meredith 1997; Sefc *et al.* 1998; Vankan and Faddy 1999). These applications of associative genetics facilitate further progress in genetic improvement through breeding. Establishment of parentage is also useful to secure legal rights of guardianship in humans, to help protect intellectual property in plant varieties, to validate breed pedigrees of domesticated animals, to protect stocks of fish, and to identify provenance of meat that is available in supermarkets (Gotz and Thaller 1998; Primmer *et al.* 2000; White *et al.* 2000).

Most studies of pedigree have utilized exclusion analysis where the molecular marker genotypes of either one or a restricted number of potential triplets of offspring and putative parents are compared. Often the identity of the mother is not in question; the maternal profile is subtracted from that of the offspring and the deduced paternal profile is then compared with candidate father genotypes (Ellstrand 1984; Hamrick and Schnabel 1985). Individuals who could not have contributed the paternal genotype are excluded; the remainder are possible parents. Nonpaternity in humans is generally declared only on the basis of exclusions exhibited by at least two unlinked and independent loci. This criterion of exclusion reduces the likelihood of a false declaration of nonpaternity on the basis of marker results that are actually due to mutation within the phylogeny. Bein *et al.* (1998) show that evidence of nonpaternity should require exclusions at loci on different chromosomes to avoid erroneous conclusions that would be made due to nondisjunction at meiosis leading to uniparental inheritance. A requirement for at least three independent exclusions to declare nonpaternity in humans has also been instituted (Gunn *et al.* 1997). In studies of natural populations of animals or plants where numerous parent-progeny triplets are examined it is usual to accept a single exclusionary event as evidence of nonpaternity (Marshall *et al.* 1998). Paternity testing has been extended to situations where DNA from either parent is unavailable. For example, paternity can still be established in circumstances where the putative father is deceased but his parents are still alive (Helminen *et al.*

[1]*Corresponding author:* Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Box 447, Houston, TX 77030-4009. E-mail: dberry@mdanderson.org

Genetics 161: 813–824 (June 2002)

Appendix D

1991; Bockel, *et al.* 1992). Chakraborty *et al.* (1994) demonstrate that paternity can be determined in cases where the mother is unavailable for testing. Lang *et al.* (1993) partially reconstructed the DNA profile of a missing crocodile parent using profiles of the mother and progeny.

Chakraborty *et al.* (1988) and Smouse and Meagher (1994) report that reliance upon exclusion alone has usually failed to unambiguously resolve paternity. Limitations have stemmed from the use of an insufficient number of independent hypervariable loci. Other statistical methods are therefore required to calculate the likelihood of paternity for each nonexcluded male (Berry and Geisser 1986; Meagher 1986; Meagher and Thompson 1986; Thompson and Meagher 1987; Devlin *et al.* 1988; Berry 1991). Marshall *et al.* (1998) draw attention to the quality of data that is encountered practically in genotypic surveys. Maternal genetic data may or may not be available, data may be absent for some candidate males, data may be missing for some loci in some individuals, null alleles exist, and typing errors occur. Reconstructing or validating the pedigrees of varieties of cultivated plants often provides additional challenges because their phylogenies can reveal apparent exclusions that masquerade as non-Mendelian inheritance. For example, apparent exclusions can occur in circumstances where an individual is used as a parent prior to completion of the inbreeding process. The development of parent and progeny then continue on parallel but separate tracks thereby allowing the possibility that alleles that are subsequently lost through inbreeding in the parent can still become fixed in the progeny. It is also possible to create many offspring from a single mating and to use the same parent repeatedly in "backcrossing." Therefore, many individual inbred lines, varieties, or hybrids can be highly related. In consequence, there are numerous (and often very similar) pedigrees. The effective number of marker loci that can discriminate between alternate pedigrees is proportionally reduced as parents are increasingly related. Consequently, inbred lines can be more similar to one or more sister or other inbreds than those inbreds are to one or both of their parents.

It has not been usual to search among hundreds of individuals to identify the most probable maternal and paternal candidates for a specific progeny. Most studies of parentage are in circumstances where there is *a priori* information for at least one of the parents (usually the maternal parent). Limited availability of marker loci and the lack of very high-throughput genotyping systems offering inexpensive datapoint costs may have focused research on studies that involve relatively few individuals and where there is at least some *a priori* indication of parentage. Studies that have been conducted without *a priori* information on parentage include species where reproductive behavior renders identification of the maternal parent difficult or impossible. Examples include those undertaken on birds that practice brood parasitism (Alderson *et al.* 1999) or extra-pair copulation (Wetton *et al.* 1992) or on species such as the wombat that are difficult to observe in the wild (Taylor *et al.* 1997).

Two circumstances favor a revised approach to the statistical analysis of pedigree. First, molecular marker technologies are rapidly developing and will allow numerous loci to be typed for thousands of individuals rapidly and inexpensively. A greater number and diversity of larger-scale studies of pedigree can be expected within the plant and animal kingdoms including individuals in which there is no prior knowledge of pedigree. A larger number of markers mean a greater chance for errors. Therefore, the second circumstance follows: Procedures that are efficient and robust in the face of apparent exclusions, missing data, and laboratory error are required.

The purpose of this article is to describe and evaluate a methodology that can be used to quantify the probability of parentage of hybrid genotypes. We focus on parentage because it is the primary focus of published literature and it is the easiest level of ancestry to understand. The method is robust in the face of mutation, pseudo-non-Mendelian inheritance (apparent exclusions) due to residual heterozygosity in parental seed sources, missing data, and laboratory error. The methodology has a number of advantages: (i) It can accommodate large datasets of possible ancestors (hundreds of inbreds or hybrids each profiled by >100 marker loci), (ii) it does not require prior knowledge about either parent of the hybrid of interest, (iii) it does not require independence of the markers, and (iv) it can successfully discriminate between many highly related and genetically similar genotypes. We demonstrate the effectiveness of this approach to identify inbred parents of maize (*Zea mays* L.) hybrids using simple sequence repeat (SSR) marker profiles for 54 maize hybrids together with their parental and grandparental genotypes included among a total of 586 inbred lines. The methodology is applicable to the investigation of parentage for all progeny developed from parental mating without subsequent generations of inbreeding.

MATERIALS AND METHODS

Algorithm: Consider an index hybrid whose parentage is unknown or in dispute. Inbreds in an available database are possible ancestors of the hybrid. The objective is to find the probabilities of closest ancestry for each inbred on the basis of information from SSRs from the index hybrid and the inbreds. There is no reason to trim the database by removing inbreds thought to be unrelated to the index hybrid because their lack of relationship will be discovered.

Consider a pair of possible ancestors, inbred i and inbred j. There is nothing special about this particular pair as all pairs will be treated similarly. The process involves calculating the probability that inbreds i and j are in the hybrid's ancestry, repeating this for all pairs of inbreds in the database.

The basis of the algorithm is Bayes' rule (*e.g.*, BERRY 1991, 1996). Let $P(i, j|\text{SSRs})$ stand for the (posterior) probability that i and j are ancestors of the index hybrid given the information from the various SSRs. Let $P(i, j)$ stand for the unconditional (or prior) probability of the same event. Finally, $P(\text{SSRs}|i, j)$ is the probability of observing the various SSR results if in fact i and j are ancestors. Bayes' rule says

$$P(i, j|\text{SSRs}) = P(\text{SSRs}|i, j) \times P(i, j)/\sum[P(\text{SSRs}|u, v) \times P(u, v)],$$

where the sum in the denominator is over all pairs of inbreds, indexed by u and v. $P(\text{SSRs}|i, j) \times P(i, j)$ is one of the terms in the denominator. (To compute the denominator in the above expression, fix a particular order to the inbreds in the database and take $u < v$ in expressions involving the pair (u, v). If there are 586 inbreds, for example, then the number of pairs and the number of terms in the denominator is $586(587)/2 = 171{,}991$.) Inbreds i and j may be parents or grandparents or other types of relations or bear no relationship at all to the hybrid. If there are more than two ancestors in the database, such as both parents and all four grandparents, then the possible pairs involving these ancestors will generally have the highest posterior probabilities. If the hybrid's true parents are in the database, then as a pair they will typically have the highest overall posterior probability. If both i and j happen to be related to one particular parent of the hybrid, then as a pair their posterior probability will be low because they will not usually account for many of the alleles that are contributed by the other parent of the hybrid.

We will make the "no-prior-information" assumption that $P(u, v)$ is the same for all pairs (u, v). This implies that this factor is cancelled from both numerator and denominator in the above expression, giving:

$$P(i, j|\text{SSRs}) = P(\text{SSRs}|i, j)/\sum P(\text{SSRs}|u, v).$$

The problem is then to calculate a typical $P(\text{SSRs}|i, j)$. Assume inbreds i and j are both ancestors. We calculate the probability of observing the resulting hybrid under this assumption. We make no assumptions about relationships among the various inbreds. Other possible ancestors will be considered implicitly in the calculation by allowing their alleles to be introduced through breedings with i and j. However, the nature of such breedings is not specified. Suppose inbred i's alleles are (a, b). Each descendant of inbred i receives one of these two alleles or not. An immediate descendant receives one with probability 1 (barring mutations). A second generation descendant receives one of them with probability 0.5. And so on. Since degree of ancestry (if any) is unknown, we label the actual probability of passing on one of these alleles to be P. Similarly, an allele from inbred j has been passed down to the hybrid or not, and the probability of the former is P. In the following, P will be taken to equal 0.50, although we will also consider $P = 0.99$ in some of the calculations.

Assuming $P = 0.50$ is consistent with the closest ancestors in the database being grandparents. However, we are not interested in grandparents *per se*. If the closest ancestors in the database were parents, then as indicated above P should equal 1 (ignoring mutations and laboratory errors). Our primary concern is when the parents are not in the database. In this case P is no greater than 0.50. Assuming $P = 0.50$ is robust over the middle range of possible values of P. One way in which it is robust is if there may be mutations and laboratory errors, in which case P would have to be < 1. Taking P to equal 0.50 levies little penalty against a particular pair in which there is an apparent exclusion from direct parentage. Therefore taking P to be < 1 means that if the true parents are in the database then they will not be ruled out if there happen to be mutations and laboratory errors. And if the closest ancestors in the database are more remote than grandparents, they are likely to be identified because they will usually have the fewest mismatches of the lines considered.

When i and j are ancestors there are four possibilities: (1) The alleles of both inbreds i and j were passed to the hybrid. (2) inbred i came through but not inbred j. (3) inbred j came through but not inbred i, and (4) neither inbred came through. Assuming independence, these have respective probabilities P^2, $P(1 - P)$, $P(1 - P)$, $(1 - P)^2$. In the case $P = 0.50$, all of these probabilities equal 0.25.

An instance of the law of total probability (Sec. 5.3, BERRY 1996) is that the probability of observing a hybrid's alleles is the average of the conditional probability of this event given the above four cases. The simplest of the four cases is the first possibility: Assuming the hybrid's alleles are passed down directly from both inbreds, the probability of observing the hybrid's genotype is either 1 or 0 depending on whether the hybrid shares both inbreds' alleles. (It is especially easy when both inbreds are homozygous.) The other three cases require an assumption regarding the possibility that an inbred's allele is not passed to the hybrid but is interrupted by a mutation, a laboratory error, or intervening breeding. We regard such an allele as being selected from all known alleles with probability 1/(number of alleles), where the number of alleles is the total number of alleles known to exist at the locus in question. An alternative approach would be to use the allelic proportions that are present in the database (or in another database). However, the lines in the database may not be randomly selected from any population. For example, a line that has been highly used in breeding would have many derivative lines in the database, in which case the frequencies of its alleles will be artificially inflated. Assuming equal probabilities for the various alleles at a given locus is robust in the sense that it is not affected by adding and dropping lines from the database.

There are many cases to consider when computing the probability of observing a hybrid's alleles, depending on the zygosity of the hybrid and the inbreds, and allowing for the possibility of missing alleles or "extra alleles" in the assessment of the hybrid and inbred genotypes. These possibilities are too numerous to list. Instead we give three simple examples. All the examples have homozygous inbreds, the most common case. And each of the three hybrids has two alleles, again the most common case. We suppose that the measured alleles for three SSRs and a particular trio of hybrid and ancestor inbreds are as we have indicated in Table 1.

For SSR 1 there are three known alleles, one in addition to alleles a and b that are listed for the three lines (hybrid, inbred i, and inbred j) in Table 1. For SSR 2 and SSR 3 there are two known alleles in addition to those listed. The calculations in the right half of Table 1 will now be explained. Implicit in calculating $P(\text{SSR}|i, j)$ is the assumption—required in both the numerator and denominator of Bayes' rule—that inbreds i and j are ancestors of the hybrid. Consider SSR 1. In case 1 above, both ancestors' alleles (as measured by the laboratory process) are assumed to pass to the index hybrid, and so in this case the hybrid is necessarily *ab*. The probability of observing the actual hybrid's genotype is 1 for case 1, as shown in Table 1. In case 2, we assume that inbred i's allele passes to the hybrid but inbred j's does not. Indeed, the hybrid has an a allele. The probability of observing a b as the other allele is 1/(number of alleles) = 1/3, as shown in Table 1. Case 3 is similar. In case 4, neither ancestor allele is passed to the hybrid; the probability of observing the hybrid's genotype (or any heterozygous genotype) is $2(1/3)(1/3) = 2/9$. Since $P = 0.50$, the overall (unconditional) probability in the rightmost column (17/36) is the simple average of the four cases, as indicated in Table 1.

For SSR 2 and SSR 3 the calculations are similar. For SSR 2 there is some evidence against pair (i, j) being ancestors,

TABLE 1

Probability of observing a hybrid's alleles using three sample SSRs and four possible combinations (cases) of alleles passed, assuming that inbreds i and j are ancestors of the hybrid

SSR	No. of alleles	Hybrid	Inbred i	Inbred j	Probability of observing the hybrid's genotype: Case 1 i, j	Case 2 i, not j	Case 3 not i, j	Case 4 not i, not j	Overall probability P(SSR|i, j)
1	3	*ab*	*aa*	*Bb*	1	1/3	1/3	2/9	17/36
2	5	*bd*	*bb*	*Cc*	0	1/5	0	2/25	7/100
3	6	*ab*	*cc*	*Dd*	0	0	0	2/36	2/144

SSR, simple sequence repeat marker profile.

but it is not conclusive. For SSR 3 there is even less evidence favoring pair (i, j). It would not take many SSRs with evidence similar to that for SSR 3 to essentially rule out this pair—provided that other pairs are not similarly inconsistent.

To find the overall P(SSRs|i, j), multiply the individual P(SSR|i, j) over the various SSRs. There are purely computational issues to address. Each P(SSR|i, j) is a number between 0 and 1. When there are a great many SSRs, the product of these numbers will be vanishingly small. To lessen problems with computational underflow, for each SSR we multiply P(SSR|u, v) by the same constant for each pair (u, v); the inverse of the largest possible such probability. For example, since 17/36 is the largest probability for a heterozygous hybrid at an SSR having three alleles (as is the case for SSR 1 in Table 1), we multiply all factors P(SSR|u, v) by 36/17. To eliminate remaining problems with underflow, we do calculations using logarithms (adding instead of multiplying) and take antilogs at the end.

The probability P(SSR|u, v) is calculated for all (u, v) pairs and summed over all possible pairings in the database, including that for the inbred pair under consideration: (i, j). This gives the denominator in the expression for P(i, j|SSRs).

To determine the probability that any particular inbred, say inbred i, is the closest ancestor of the index hybrid, sum P(SSR|i, v) over all inbreds v with $v \neq i$. Call this P(i|SSRs). The maximum of P(i|SSRs) for any inbred i is 1. But since there is one closest ancestor on each side of the family, the sum of P(i|SSRs) over all inbreds i is 2. If there is a particular pair (i, j) for which P(i, j|SSRs) is close to 1 then both P(i|SSRs) and P(j|SSRs) separately will be close to 1.

SSR data: DNA was extracted from 34 maize hybrids and from 586 maize inbreds. All of the hybrids and most inbreds are proprietary products of Pioneer Hi-Bred International; some important publicly bred inbred lines were also included. The inbred parents and grandparents of each hybrid were included within the set of inbreds. Other inbreds that were genotyped include many that are highly related by pedigree to parents and grandparents of the hybrids. The hybrids were chosen because each has a pedigree that is known to us and collectively they represent a broad array of diversity of maize germplasm that is currently grown in the United States ranging from early to late maturity.

A total of 195 SSR loci were used in this study following procedures described in Smith *et al.* (1997), but modified as described below. SSR loci were chosen on the basis that they individually have been shown to have a high power of discrimination among maize inbred lines and collectively they provide for a sampling of diversity for each chromosome arm. Of these SSR loci, the following numbers (in parentheses) were located on individual maize chromosomes as follows: 1 (35), 2 (26), 3 (22), 4 (20), 5 (16), 6 (9), 7 (6), 8 (18), 9 (12), and 10 (14); 17 SSR loci have not yet been mapped. The correlations among the loci are unknown and are irrelevant for our methodology.

Sequence data for primers that allow many of these (and other) SSR loci to be assayed are available at website http://www.agron.missouri.edu. All primers were designed to anneal and amplify under a single set of conditions for PCR in 10-μl reactions. Genomic DNA (10 ng) was amplified in 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris-Cl (pH 8.3) using 0.3 units AmpliTaq Gold DNA polymerase (PE Corporation) oligonucleotide primer pairs (one primer of each pair was fluorescently labeled) at 0.17 μM and 0.2 mM dNTPs. This mixture was incubated at 95° for 10 min (hot start); amplified using 45 cycles of denaturation at 95° for 50 sec, annealing at 60° for 50 sec, extension at 72° for 85 sec; and then terminated at 72° for 10 min. A water bath thermocycler manufactured at Pioneer Hi-Bred International was used for PCR reactions. PCR products were prepared for electrophoresis by diluting 3 μl of each product to a total of 27 μl using a combination of PCR products generated from other loci for that same maize genotype (multiplexing) and/or dH20. Dilution of 1.5 μl of this mixture to 5 μl with gel loading dye was performed; it was then electrophoresed at 1700 V for 1.5 hr on an ABI model 377 automated DNA sequencer equipped with GENESCAN software v. 3.0 (PE-Applied Biosystems, Foster City, CA).

PCR products were sized automatically using the "local Southern" sizing algorithm (Elder and Southern 1987). After sizing of PCR products using GeneScan, alleles were assigned using Genotyper software (PE-Applied Biosystems). Generally, allele assignations for each locus were made on the basis of histogram plots consisting of 0.5-bp bins. Breaks between the histogram plots of >1 bp were generally considered to constitute separation between allele bins; however, other criteria, such as the presence of the nontemplate-directed addition of adenine (+A addition) and naturally occurring 1-bp alleles, were used on a marker-by-marker basis to define the allele dictionary. All allele scores were made without knowing the identities of the maize genotypes.

RESULTS

Table 2 presents the probability of closest ancestry of the top five ranking inbred lines for each of 5 hybrids at $P = 0.50$ (Table 2A) and $P = 0.99$ (Table 2B). Probabilities of ancestry are shown for all 34 hybrids and the top ranking inbreds in Figure 1: $P = 0.50$ (Figure 1a) and $P = 0.99$ (Figure 1b). Results for the hybrids presented in Table 2 are featured at the top of Figure 1.

TABLE 2

Probability of ancestry of five hybrids using data obtained from 50, 100, and 195 SSR loci

	50 loci			100 loci			195 loci		
Hybd.	Inbd.	Prob.	SE	Inbd.	Prob.	SE	Inbd.	Prob.	SE
				A. Assuming $P = 0.50$					
3417	SP1	0.9607	0.0125	P1	0.8749	0.0232	P1	1.0000	E-07
	P2	0.8077	0.1965	P2	0.8141	0.2235	P2	0.9957	0.0033
	D2P2	0.1016	0.1038	D1P2	0.1859	0.2235	D1P2	0.0043	0.0033
	D1P2	0.0907	0.0927	SP1	0.1243	0.025	D2P2	E-06	E-06
	P1	0.082	0.0125	D1P1	0.0009	0.0002	SP1	E-06	E-07
3525	P1	0.8545	E-07	P1	0.9999	<E-20	P1	1.0000	<E-20
	P2	0.8188	E-07	P2	0.5437	<E-20	P2	0.9635	0.0528
	D1P2	0.1699	E-07	D1P2	0.4563	<E-20	D1P2	0.0365	0.0528
	GP1	0.1441	E-07	GP1	E-07	E-18	SP1	E-15	<E-20
	GP2	0.0110	E-08	SP1	E-07	<E-20	GP2	E-16	<E-20
3556	P1	1.0000	E-06	P1	0.9999	E-10	P1	1.0000	<E-20
	P2	0.9616	E-08	P2	0.9997	E-10	P2	1.0000	<E-20
	D1P2	0.0340	E-10	D1P2	0.0003	E-14	D1P2	E-09	<E-20
	GP2	0.0043	E-09	D2P2	E-05	E-15	D2P2	E-14	<E-20
	D2P2	0.0002	E-10	D3P2	E-06	E-17	GGP2	E-17	E-17
3905	D1P1	0.9822	E-08	D1P1	0.9803	0.0058	P1	1.0000	E-08
	SP2	0.4927	E-07	SP2	0.6280	0.0976	D1P2	1.0000	E-06
	D2P2	0.2886	E-07	D1P2	0.2321	0.0617	D2P2	E-06	E-06
	D1P2	0.1622	E-07	D2P2	0.1317	0.0372	P2	E-07	E-13
	P2	0.0565	E-07	P1	0.0197	0.0058	D3P2	E-10	E-16
3940	P2	0.9997	0.0001	P2	0.9999	E-05	P2	1.0000	E-09
	D1P2	0.9203	0.0009	P1	0.9970	0.0011	P1	1.0000	E-09
	P1	0.0648	E-05	D1P2	0.0030	0.0011	D1P2	E-11	E-11
	D1P1	0.0127	E-05	D2P2	0.0001	E-05	DP1P2	E-17	E-17
	DP1P2	0.0014	0.0009	DP1P2	0.0001	E-07	D2P2	E-19	E-18
				B. Assuming $P = 0.99$					
3417	SP1	0.9995	0.0001	P1	0.9999	E-05	P1	0.9999	E-08
	P2	0.8836	0.1658	P2	0.9938	0.0107	P2	0.9999	E-08
	D1P2	0.0722	0.1029	D1P2	0.0061	0.0107	D1P2	E-11	E-11
	D2P2	0.0441	0.0628	D1P1	E-05	E-06	D2P2	E-14	E-14
	P1	0.0004	0.0001	SP1	E-05	0	SP1	E-20	E-21
3525	P1	0.9999	0	P1	0.9999	0	P1	1.0000	0
	P2	0.8991	0	D1P2	0.9749	0	P2	0.6135	0.4446
	D1P2	0.1008	E-11	P2	0.025	0	D1P2	0.3864	0.4446
	GP1	E-05	0	D2P2	E-20	0	GP2	E-48	0
	GP2	E-06	E-17	SP1	E-24	0	D2P2	E-49	0
3556	P1	1.0000	0	P1	1.0000	0	P1	0.9999	0
	P2	0.9996	0	P2	0.9999	0	P2	0.9999	0
	D1P2	0.0003	0	D1P2	E-09	0	D1P2	E-22	0
	D1P1	E-11	0	D3P1	E-21	0	D2P1	E-49	0
	D2P1	E-13	0	D2P1	E-21	0	D3P1	E-54	0
3905	D1P1	0.9999	0	D1P1	0.9999	E-08	P1	1.0000	E-09
	P2	0.9992	0	P2	0.9999	E-06	P2	0.9947	E-09
	SP2	0.0006	0	D1P2	E-06	E-06	D1P2	0.0052	E-11
	D1P2	E-05	0	SP2	E-07	E-13	D2P2	E-18	E-18
	D2P2	E-06	0	D2P2	E-09	E-10	D1P1	E-25	E-25
3940	P2	0.9999	E-08	P2	1.0000	E-08	P1	1.0000	E-09
	D1P2	0.9999	E-08	P1	0.9999	E-05	P2	1.0000	E-09
	P1	E-06	E-13	D1P2	E-05	E-05	D1P2	E-24	E-24
	D1P1	E-08	E-15	D2P2	E-12	E-11	DP1P2	E-44	E-44
	DP1P2	E-12	E-[illegible]	DP1P2	E-21	E-21	D2P2	E-50	E-49

Hybd., hybrid; Inbd., inbred; Prob., probability; SE, standard error, referring to the variability in the results of the runs; P1, parent one; P2, parent two; SP1, SP2, full sibling of parent one/parent two; D/P1/D/P2, derivatives of parent one/parent two, index *i* for distinct inbred lines; DP1P2, derivatives of both parent one and parent two.

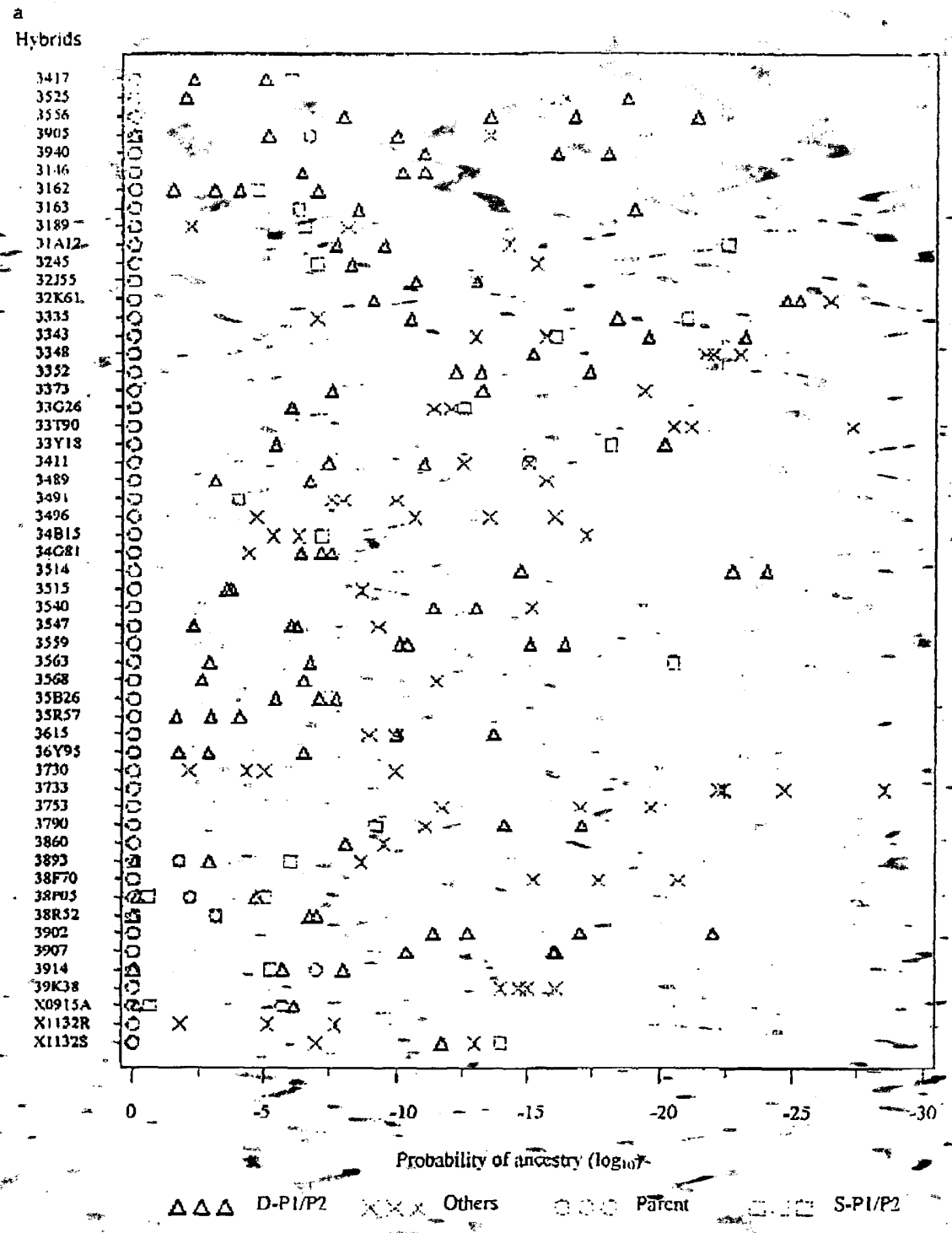

FIGURE 1.—(a) Probabilities of ancestry, assuming $P = 0.50$, for all 54 hybrids and top ranking inbreds—those with probability of ancestry at least 10^{-6}. (b) Probabilities of ancestry, assuming $P = 0.99$, for all 54 hybrids and top ranking inbreds—those with probability of ancestry at least 10^{-6}.

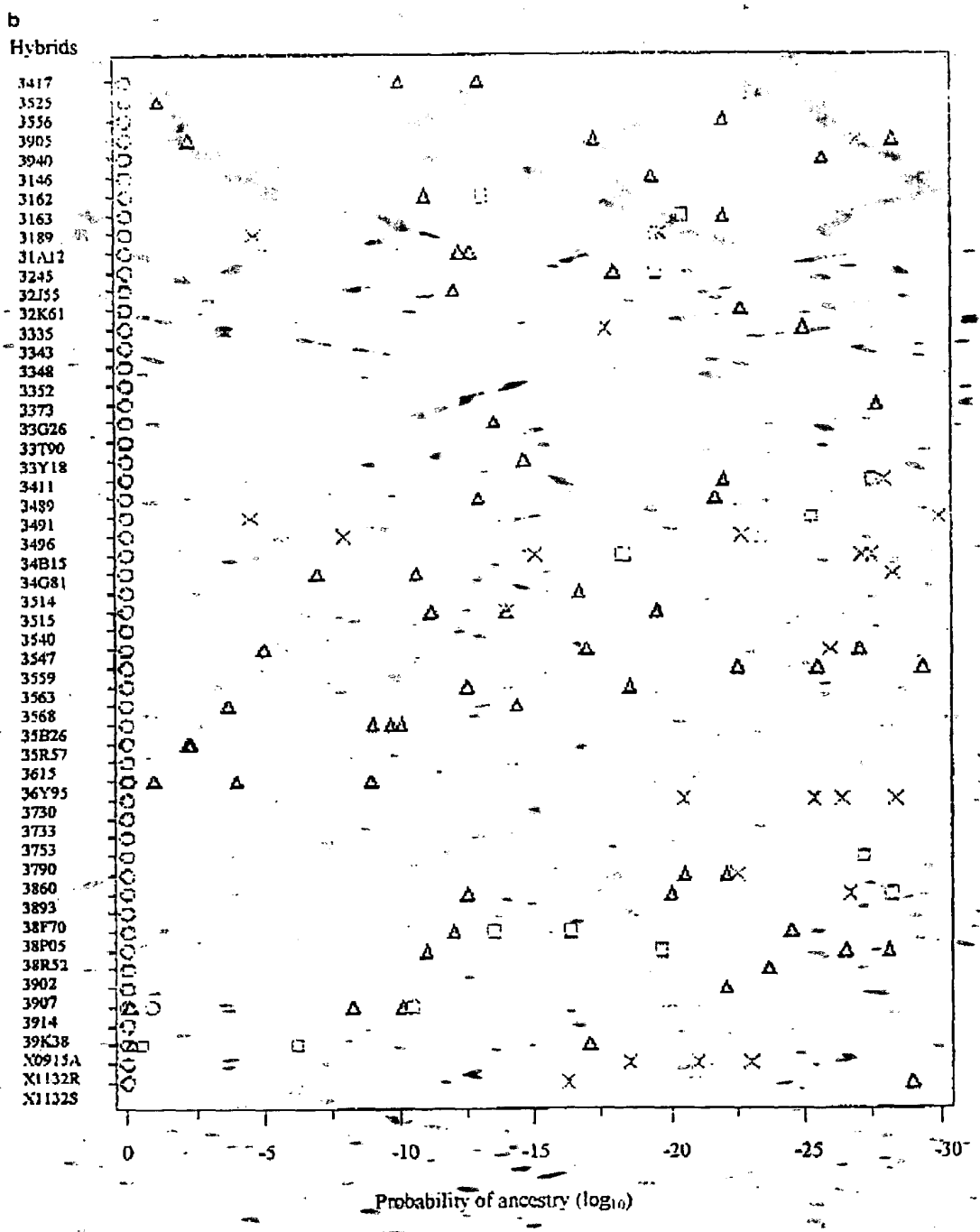

FIGURE 1.—*Continued.*

When the algorithm used $P = 0.50$, the two correct parents were identified as highest in probability for 48 (89%) hybrids (Figure 1). For each of 6 hybrids (3893, 38P05, 38R52, 3905, 3914, and X0915A), one parent ranked in the top two places. The other parent was supplanted either by a sister inbred or by an inbred that was a direct progeny of that parent. Overall, 102 (94%) of 108 parental inbreds were correctly identified. For hybrids where both parents ranked first or second, the range of probabilities for parental lines that ranked first from among all other inbreds ranged from 1.0000 to 0.9997; parental lines ranking second ranged from

1.0000 to 0.9653. For 35 hybrids, both parents had probabilities of ancestry in excess of 0.999. Probabilities of ancestry for nonparents that ranked in first or second places were from 0.9999 to 0.7054. For the majority of hybrids, the probability of the third and highest ranked nonparental inbred was at or below E-06. This indicates that there is usually very little uncertainty about closest ancestors.

When the algorithm used $P = 0.99$ to examine each of the 54 hybrids, both parents were correctly identified for 52 (96%) of hybrids and for 98% (102/104) of the parents across all hybrids (Figure 1). Two hybrids (3914 and X0915A), in which one parent was not ranked in the top two, were also in the subset not ranked in the top two assuming $P = 0.50$ (above). In both cases their ranks improved (both to third rank) and the actual parent was supplanted by an inbred that was a direct progeny of the corresponding parental line. For 49 hybrids, both parents had probabilities of ancestry in excess of 0.999. Among the 5 hybrids having a parent ranking second with a probability of ancestry below 0.999, the lowest of these probabilities was 0.8976 and the highest probability for a third ranking nonparent was 0.1023. For most hybrids the probability for the third and highest ranked nonparental inbred was at or below E-10.

Table 2 also addresses data analysis in circumstances where heterozygous loci occur in inbred lines or where a hybrid is scored for the presence of more than two alleles per locus. The presence of more than a single allele per locus in inbred lines is an infrequent occurrence in well-maintained inbred development and seed increase programs but is possible because ~3–5% of loci can still be segregating and unintended pollination from genotypes not designated as parents of the hybrid can occur. For hybrids, more than two alleles per locus can be scored when DNA is extracted from a bulk of individual plants and because inbred parents are not homozygous due either to residual heterozygosity or to contamination or because one or more direct parents of the hybrid are themselves hybrids. The presence of more than one allele per locus in an inbred line and more than two alleles per locus in a hybrid therefore can be accommodated by multiple runs of the algorithm, each with a random choice of two alleles per locus. Consequently, standard errors in the case of analyzing data from 195 loci tend to be very small because there were few loci where an inbred or hybrid sample (from a bulk of individual plants) was scored for more than two alleles.

Marshall *et al.* (1998) have drawn attention to errors that can be encountered in genotyping surveys. These errors include missing data, null alleles, and typing errors. We therefore investigated the robustness of the algorithm by examining the effects of modifications in the data for five hybrids (3417, 3525, 3556, 3905, and 3940). First, we reduced the number of SSRs used, from the full set of 195 to 100 and then to 50 (Table 2). Use of 50 loci generated incorrect rankings of one parent for each of two hybrids (3417 and 3940) and for both parents of one hybrid (3905). All of these most highly ranked nonparental inbreds were closely related to the true parents for each of the respective hybrids; six different inbred lines were involved. Four were direct progeny of the true parents (one with additional backcrosses from the true parent) and two were full sisters (from a cross of highly related inbreds) of the actual parent of the hybrid. Using 100 loci resulted in correct parental rankings for all hybrids except for 3905 where neither parent ranked in first or second place. Four inbreds outranked the true parents of 3905. All four nonparents were closely related to the respective true parents; three were direct progeny of the true parent of the hybrid (one with additional backcrossing to that parent) and one was a full sister of the true parent. Use of data from all 195 loci corrected the placement for one of the parents of hybrid 3905. Two inbreds that were not parents of this hybrid remained ranked more highly than one of the true parents. Both were direct progeny of that parent, and one of these inbreds had additional backcrossing to that parent in its pedigree.

To address the consequences of laboratory and other sources of error, we artificially compromised data quality beyond the level originally provided by eliminating specific proportions of alleles that had been scored (establishing scenarios where various numbers of SSR alleles were not scored) and by misscoring other alleles (establishing scenarios where various numbers of SSR alleles were scored incorrectly). We also combined the scenarios of missing data and wrongly scored data. Table 3 contains a summary of the results of making these modifications in the data. For all modifications we used data from all SSR loci and we also randomly chose SSR loci to create subsets of 50 and 100 loci. In each case, the program was run 20 times for each hybrid/set of loci. When all 195 loci were examined, replications differed only according to the particular choice of alleles for loci where more than two alleles had been scored.

To evaluate robustness in the face of missing data or mistyped data, we simulated individual and combined categories of these data in the hybrid and all inbred lines at levels of 2, 5, 10, and 25% of the alleles for each of five hybrids and all inbreds beyond the level of error as originally scored by the laboratory. We examined the effects of these levels and types of error for three sizes of database: 50 loci, 100 loci, and all 195 scored loci. The same five hybrids considered in Table 2 were investigated: 3417, 3525, 3556, 3905, and 3940. One of these hybrids (3905) was chosen because one of its parents did not rank among the top two places even when the complete and unmodified data from all SSR loci were used.

Examples of robustness in the face of additional error

TABLE 3

Number of parents ranked in first and second positions (maximum is 2)

Type of simulated data	% level change	No. of loci:	Hybrid 3417 50	3417 100	3417 195	3525 50	3525 100	3525 195	3556 50	3556 100	3556 195	3905 50	3905 100	3905 195	3940 50	3940 100	3940 195	Mean % max.
Missing	0		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
	2		1	2	2	2	1	2	2	2	2	0	0	1	1	2	2	73
	5		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
	10		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
	25		0	2	2	1	1	1	2	2	2	0	0	1	1	1	1	57
	Mean % max.		40	100	100	90	80	90	100	100	100	0	0	50	50	90	90	
Miscored	0		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
	2		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
	5		1	2	2	2	1	2	2	2	2	0	0	1	1	2	2	73
	10		1	1	2	2	2	2	2	2	2	0	0	1	1	2	2	73
	25		1	0	2	1	1	2	1	2	2	0	1	1	1	2	2	63
	Mean % max.		50	70	100	90	80	100	90	100	100	0	10	50	50	100	100	
Missing plus miscored	0		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
	2		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
	5		1	2	2	2	1	1	2	2	2	0	0	1	1	2	2	70
	10		1	2	2	1	2	2	2	2	2	0	1	1	1	2	2	77
	25		0	1	2	0	0	0	0	0	0	0	0	1	0	0	1	17
	Mean % max.		40	90	100	70	70	70	80	80	80	0	10	50	40	80	90	
Overall mean			43	87	100	83	77	87	90	93	93	0	7	50	47	90	93	

Hybrids considered are the same as those in Table 2.

for five hybrids using subsets of 50 and 100 loci and all loci are shown in Table 9 where numbers of parents ranking into the top two places are presented. Degradation in the preferential ranking of parent inbreds at a level of 25% additional missing data was shown for one hybrid (3525) with usage of 50, 100, or all SSR loci. Degradation in the preferential ranking of parent inbreds at a level of 25% additional misscored data was shown for hybrid 3556. When both additional levels of missing and misscored data were simulated, degradation in the ability to preferentially rank inbred parents occurred for all hybrids and for all sets of SSR (50, 100, and 195 loci) except for hybrid 3417 when data from 195 SSR loci were used. Over all five hybrids, use of 100 loci improved robustness from the use of 50 loci; use of 195 loci further improved robustness for four hybrids (3417, 3525, 3905, and 3940). The degree of improvement was small, except for hybrid 3905.

We also ranked inbreds according to their probability of ancestry of hybrids when both parents and all inbred derivatives and full-sister inbreds of the respective inbred parents for each hybrid were excluded from the analysis. The results are too voluminous to present here but can be summarized as follows: Using $P = 0.50$, a grandparent of each respective hybrid ranked into first place for 41 (76%) hybrids; probabilities ranged from 0.4976 to 1.0 and most were above 0.9999. Other classes of inbreds that ranked in first position for probability of ancestry were inbreds derived directly by pedigree from a grandparent of the respective hybrid (DGP) for 13% of hybrids, inbreds derived directly by pedigree from a great-grandparent of the respective hybrid (DGGP) for 9% of hybrids, and one class (2% of hybrids) with an inbred ranked into first place that was directly related by pedigree to the great-great-grandparent of that hybrid. Inbreds that ranked in second position were related to the respective parents of the hybrid as follows: Thirty-one (57% of hybrids) were a grandparent of the respective hybrid, 11 (20%) were classed as DGP, 7 (13%) were DGGP, 1 (2%) was class DGGGP, and 4 (7%) were a great-grandparent (GGP) of the respective hybrid. Over all hybrids, two of the four grandparents ranked into first and second positions for 23 (43% of hybrids); three grandparents ranked into the first three positions for 5 (9% of hybrids). There were no instances where all four grandparents ranked into the first four positions. Thirty hybrids had a grandparent ranked into first position using $P = 0.99$. The number of grandparents ranked into the top five positions was 93 (compared to 108 when $P = 0.50$). The number of grandparents ranking into the top two positions was 55 (compared to 71 when $P = 0.50$). The mean probability of a grandparent that ranked into the first two positions was 0.9288 (SD = 0.1454) when $P = 0.50$ and 0.9980 (SD = 0.0104) when $P = 0.99$.

DISCUSSION

The prevalent use of paternity indices demonstrates that it is advantageous to have explicit probabilities of ancestry to distinguish among different pedigrees. Molecular marker profiles are rapidly becoming more extensive and cost effective to generate. Features that would advance the statistical analysis of molecular marker data to provide explicit probabilities of ancestry include the ability to calculate probabilities of ancestry where there is no *a priori* information as to the identity of one (usually the maternal) parent and robustness in the face of laboratory error.

Maize inbred lines and hybrids provide a very exacting set of materials for evaluating the discriminatory abilities of molecular data and statistical procedures that are employed to interpret those data. Hundreds of maize inbred lines of known pedigree together encompass a great diversity and complexity of pedigree relationships. Some inbred lines can be very highly related and genetically similar due to their derivation from common parentage including from parents that are themselves highly related. Consequently, relationship categories such as "sister" or "parent" when applied to maize inbreds usually refer to closer degrees of pedigree relationship and, thus, of germplasm and molecular marker profile similarity than those of the equivalently named classes of relationship for animal species. Most maize hybrids that are widely used in the United States today are constructed from pairs of inbred lines that are unrelated by pedigree, each inbred parent having been bred from a separate "pool" of germplasm. Various degrees of relatedness are possible between hybrids according to the pedigree relationships among their constituent inbred parents.

Using $P = 0.99$ in the algorithm is more specific for identifying parents than using $P = 0.50$. However, $P = 0.99$ is less robust for identifying other relatives, such as grandparents. When the algorithm was run at $P = 0.50$ there were 6 hybrids for which one parent did not rank among the top two most probable genotypes. For the remaining 48 hybrids the correct parents were identified even in circumstances where other candidate inbreds included not only full-sister lines bred from related parents but also inbreds even more closely related to the true parent by virtue of being backcross conversions of the inbred parent of the hybrid. For each of the 6 hybrids where a nonparent ranked above a true parent, that higher ranked inbred was always either a sister or progeny of the outranked true parent. The range of pedigree relationships as expressed by the Malécot coefficient of relatedness (MALÉCOT 1948) that was encompassed by pairs of true parents and more highly ranked inbred relatives of the true parents was from 0.8390 to 0.9680. A coefficient of 0.8390 approximates a relationship between inbred A and A′ where

inbred A' has been bred from a cross of inbreds A and B with between one and two additional backcrosses of the parental inbred A. A Malécot coefficient of relationship of 0.9680 closely approximates a relationship between inbreds A and A" where four additional backcrosses of parental inbred A follow the initial cross of inbreds A and B.

Running the algorithm at $P = 0.99$ in comparison to $P = 0.50$ raises the probability of ancestry for the parents while diminishing the probabilities for the third and lower ranking candidate inbred lines. Use of the algorithm at $P = 0.99$ increased both the percentage of hybrids with both parents ranked in the first two positions (from 89 to 96%) and the percentage of parental inbreds that were ranked first and second (from 94 to 98%). Two hybrids (3914 and X0915A) did not have both parents ranked first and second when the algorithm was run at $P = 0.99$. For both of these hybrids the nonparental inbred that outranked the true parent was itself a product by pedigree from the true parent that had been created by an additional four backcrosses of that parent; the Malécot coefficient of relationship between the parent of the hybrid and the inbred that outranked that parent for these two hybrids was 0.9636.

Robustness was tested by evaluating the effects of using data from different numbers of loci and by simulating additional levels of missing and misscored data up to combined levels of 25% error beyond that which was provided by the laboratory. From our experience, error rates of 5 to 10% can occur in SSR profiling of maize due chiefly to the combined effects of residual heterozygosity among seed lots and by deficiencies in the scoring of heterozygotes in hybrids. The additional levels of simulated error, therefore, include values (up to ~35% total error) that are well outside of our experience. For five hybrids that were examined, increasing the number of loci from 50 to 100 (with no additional missing or misscored data) did reduce the number of instances where inbreds that were not parents of a hybrid outranked the true parent from four to one. Nonetheless, all of these more highly ranked inbreds, although they were not themselves the true parents of the respective hybrid, were either direct progeny or full sisters of the true parent (Table 2). Consequently, if such degrees of error can be tolerated in respect of pedigrees for inbreds that are identified as parents of hybrids, then SSR data from 50 loci of equivalent discrimination ability are sufficient. Use of data from 50 loci also evidenced robustness in the face of up to 10% additional levels of either missing or misscored data; no degradation in the ability to identify a parent was apparent up to the level of 10% additional error except for 10% additional missing and misscored alleles for one hybrid (3525; Table 3). However, use of 100 loci increased the proportion of true parents that were correctly identified from 53% (for 50 loci) to 71% (mean correct parents over all levels of error; Table 3). Use of data from 195 loci provided greater resiliency against additional levels of error. However, use of data from 195 loci was unable to provide resiliency against the negative effects of adding combined levels (at 25%) of both missing and misscored data (Table 3). At the 25% level of additional poor data integrity, inbreds that were not related to the true parent of the hybrid outranked the true parent for four of the five hybrids. Levels of missing or misscored data should, therefore, be kept below 15–20% (assuming a level of 5–10% error in the data we analyzed prior to simulating additional error).

We have previously examined the pedigrees of inbreds that are ranked into the first two positions when the true parents are removed from the list of candidate inbred lines. Usually, direct progeny or full sisters of the true parents then rank most highly (data not presented). We therefore examined the rankings of inbreds with respect to their ranking and probability of inclusion in the ancestry of each hybrid after the removal, not only of the true parents, but also of the progeny of the true parents and any full sisters of the true parents. In these circumstances the grandparents of the hybrids are ranked predominantly into top positions. Using $P = 0.50$, a grandparent ranked into first position for 76% hybrids and into second position for 57% hybrids; with $P = 0.99$ a grandparent ranked into first place in 56% of hybrids. At $P = 0.50$ two grandparents ranked into first and second positions for 43% hybrids and into the first three positions for an additional 9% hybrids. Most of the remaining inbreds that ranked into the top two positions were progeny of the grandparent. A total of 108 grandparents ranked into the top five positions when $P = 0.50$; 93 ranked into these positions when $P = 0.99$. Seventy-one grandparents ranked into the top two positions when $P = 0.50$; 55 grandparents ranked into these positions when $P = 0.99$. The mean probability of a grandparent in the top two positions was 0.9288 (SD 0.1454) when $P = 0.50$ and 0.9980 (SD 0.0104) when $P = 0.99$. Our algorithm was written to identify pairs of ancestors; alternative algorithms could be tailored to identify all grandparents once parents had been identified and removed from the list of candidate inbreds.

We have demonstrated the capability and robustness of an algorithm that can be used to show probability of parentage in circumstances where the *a priori* pedigree identity of neither parent is known. Exclusions are taken into account, thereby allowing parentage to be shown even when the two parents are not represented in the database of molecular profiles that are examined. Heterozygous candidate parents can be accommodated. The number of loci that is necessary to provide a reliable basis of determining pedigree is dependent upon the degree of relatedness among parents and nonparents and upon the discriminatory ability of the marker system

in the species of interest. Using $P = 0.99$ compared to $P = 0.50$ preferentially identified more true parents and with a greater difference of probability to third placed nonparents. If there is reasonable assurance that the parents are among the candidate list of inbreds, then $P = 0.99$ should be used; if greater robustness is required, then $P = 0.50$ should be used.

Applications of our algorithm include the identification of pedigrees among individuals of plant or animal species where molecular profile datasets exist that can be interpreted in terms of segregating alleles at individual marker loci and that provide a sufficient power of discrimination. Capabilities to generate large datasets of suitable molecular profile data are already available and are increasing rapidly with the advent of single nucleotide polymorphisms. One further application of our algorithm is to assist in the protection of intellectual property that is obtained on plant varieties or upon specific dams or sires of animals through the determination of pedigrees.

LITERATURE CITED

Alderson, G. W., H. L. Gibbs and S. G. Sealy, 1999 Parentage and kinship studies in an obligate brood parasitic bird, the brown-headed cowbird (*Molothrus ater*), using microsatellite DNA markers. J. Hered. 90: 182–190.

Bein, G., B. Driller, M. Schürmann, P. M. Schneider and H. Kirchner, 1998 Pseudo-exclusion from paternity due to maternal uniparental disomy 16. Int. J. Leg. Med. 111: 328–330.

Berry, D. A., 1991 Inferences using DNA profiling in forensic identification and paternity cases (with discussion). Stat. Sci. 6: 175–205.

Berry, D. A., 1996 *Statistics: A Bayesian Perspective*. Duxbury Press, Belmont, CA.

Berry, D. A., and S. Geisser, 1986 Inferences in cases of disputed paternity, pp. 353–382 in *Statistics and the Law*, edited by M. H. DeGroot, S. E. Fienberg and J. K. Kadane. Wiley Publishing, New York.

Bockel, B., P. Nürnberg and M. Krawczak, 1992 Likelihoods of multilocus DNA fingerprints in extended families. Am. J. Hum. Genet. 51: 554–561.

Bowers, J. E. and C. P. Meredith, 1997 The parentage of a classic wine grape, Cabernet Sauvignon. Nat. Genet. 16: 84–87.

Chakraborty, R., T. R. Meagher and P. E. Smouse, 1988 Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genetics 118: 527–536.

Chakraborty, R., L. Jin and Y. Zhong, 1994 Paternity evaluation in cases lacking a mother and nondetectable alleles. Int. J. Leg. Med. 107: 127–131.

Devlin, B., K. Roeder and N. C. Ellstrand, 1988 Fractional paternity assignment: theoretical development and comparison to other methods. Theor. Appl. Genet. 76: 369–380.

Elder, J. K., and E. M. Southern, 1987 Computer-aided analysis of one dimensional restriction fragment gels, pp. 165–172 in *Nucleic Acid and Protein Sequence Analysis—A Practical Approach*, edited by M. J. Bishop and C. J. Rawlings. IRL Press, Oxford.

Ellstrand, N. C., 1984 Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. Am. Nat. 123: 819–828.

Goltz, K., and G. Thaller, 1998 Assignment of individuals to populations using microsatellites. J. Anim. Breed. Genet. 115: 53–61.

Gunn, P. R., K. Trueman, P. Stapleton and D. B. Kearkovski, 1997 DNA analysis in disputed parentage: the occurrence of two apparently false exclusions of paternity, both at short tandem repeat (STR) loci, in the one child. Electrophoresis 18: 1650–1652.

Hamrick, J. L., and A. Schnabel, 1985 Understanding the genetic structure of plant populations: some old problems and a new approach, pp. 30–50 in *Population Genetics in Forestry*, edited by H. R. Gregorius. Springer-Verlag, Heidelberg, Germany.

Helminen, P. V. Johnsson, C. Ehnholm and L. Peltonen, 1991 Proving paternity of children with deceased fathers. Hum. Genet. 87: 657–660.

Lang, J. W., R. K. Aggarwal, K.-C. Majumdar and L. Singh, 1993 Individualization and estimation of relatedness in crocodilians by DNA fingerprinting with a 33mer-derived probe. Mol. Gen. Genet. 238(1–2): 49–56.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson et Cie, Paris.

Marshall, T. C., J. Slate, L. E. B. Kruuk and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7: 639–655.

Meagher, T. R., 1986 Analysis of paternity within a natural population of *Chamaelirium luteum* (L.). Identification of most-likely male parents. Am. Nat. 128: 199–215.

Meagher, T. R., and E. Thompson, 1986 The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. Theor. Popul. Biol. 29: 87–106.

Miller, P. S., 1975 Selective breeding programs for rare alleles: examples from the Przewalski's horse and California Condor pedigrees. Conserv. Biol. 9: 1262–1273.

Primmer, C. R., M. T. Koskinen and J. Piironen, 2000 The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proc. R. Soc. Lond. B Biol. Sci. 267: 1699–1704.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94: 9197–9201.

Sefc, K. M., H. Steinkellner, J. Glossl, S. Kampfer and F. Regner, 1998 Reconstruction of a grapevine pedigree by microsatellite analysis. Theor. Appl. Genet. 97: 227–231.

Smith, J. S. C., E. C. L. Chin, H. Shu, O. S. Smith, S. J. Wall *et al.*, 1997 An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPs and pedigree. Theor. Appl. Genet. 95: 163–173.

Smouse, P. E., and T. R. Meagher, 1994 Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) gray (Liliaceae). Genetics 136: 313–322.

Taylor, A. C., A. Horsup, C. N. Johnson, P. Sunnucks and B. Sherwin, 1997 Relatedness structure detected by microsatellite analysis and attempted pedigree reconstruction in an endangered marsupial, the northern hairy-nosed wombat *Lasiorhinus krefftii*. Mol. Ecol. 6: 9–19.

Thompson, E., and T. R. Meagher, 1987 Parental and sib likelihoods in genealogy reconstruction. Biometrics 43: 585–600.

Vaneau, D. M., and M. J. Faddy, 1999 Estimations of the efficacy and reliability of paternity assignments from DNA microsatellite analysis of multiple-sire matings. Anim. Genet. 30: 355–361.

Wetton, J. H., D. T. Parkin and R. E. Carter, 1992 The use of genetic markers for parentage analysis in *Passer domesticus* (house sparrows). Heredity 69: 243–251.

White, E., J. Hunter, C. Dubetz, R. Brost, A. Bratton *et al.*, 2000 Microsatellite markers for individual tree genotyping: application in forest crime prosecutions. J. Chem. Technol. Biotechnol. 75: 923–926.

Communicating editor: Z-B. Zeng